

Conference Proceedings

Edited by:
Luciana Duranti and
Elizabeth Shaffer

The Memory of the World in the Digital Age: Digitization and Preservation

An international conference
on permanent access to
digital documentary heritage

Hosted by:



a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

In collaboration with



UNIVERSITY OF
TORONTO



United Nations
Educational, Scientific and
Cultural Organization



Memory of the World
20th Anniversary

26 to 28 SEPTEMBER 2012

Vancouver, British Columbia, Canada
Sheraton Vancouver Wall Centre



Conference Proceedings

Edited by:
Luciana Duranti and
Elizabeth Shaffer

The Memory of the World in the Digital Age: Digitization and Preservation

An international conference
on permanent access to
digital documentary heritage

26 to 28 SEPTEMBER 2012
Vancouver, British Columbia, Canada
Sheraton Vancouver Wall Centre

UNESCO Memory of the World Programme, Knowledge Societies Division

This book of Proceedings includes most of the papers and posters presented at the International Conference “The Memory of the World in the Digital Age: Digitization and Preservation” held on 26–28 September 2012 in Vancouver, British Columbia, Canada, by the UNESCO Memory of the World Programme, Knowledge Societies Division, and The University of British Columbia in collaboration with the University of Toronto.

The proceedings have been compiled and formatted with minor editing; papers and posters appear as submitted. The authors are responsible for the choice and the presentation of the facts contained in this publication and for the opinions they express, which are not necessarily those of UNESCO and do not commit the Organization.

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The organizers of this UNESCO Memory of the World Programme Conference would like to sincerely thank everyone who contributed to the Conference in Vancouver and to these proceedings.

Published by UNESCO 2013, with the financial support of the Social Sciences and Humanities Research Council of Canada | Conseil de recherches en sciences humaines du Canada (SSHRC) and the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) Project.



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada



InterPARES Project

International Research on Permanent Authentic Records in Electronic Systems

Preface

This publication presents the proceedings of the international conference 'Memory of the World in the Digital Age: Digitization and Preservation' which was held in Vancouver, Canada, from 26 to 28 September 2012.

More than 500 experts and other interested persons from all regions of the world participated in this knowledge-sharing and policy-driving event to discuss and exchange opinions on how to protect the world's documentary heritage. Although this heritage is the record of knowledge, its physical carriers are extremely vulnerable and can easily disappear without a trace. Whether recorded on a clay tablet or an electronic tablet, our methods of sharing content and knowledge need to be protected.

It is impossible to exaggerate the importance of documentary heritage in our lives. It governs our actions whether these relate to creating the basis of mutual respect between different civilizations and communities or building knowledge societies. Documentary heritage provides the foundation of peace, our identity and knowledge.

UNESCO's interest in this subject matter is as fundamental as its constitution with its mandate to contribute to building peace through the spread of knowledge from improved access to printed and published materials. These core materials, our documentary heritage, have been preserved in archives, libraries and museums for generations.

But while measures needed to maintain access to print materials are globally understood, the newer challenges related to preserving digital information are not keeping pace with technological development. The need for dedicated hardware and software, associated with their rapid obsolescence, hamper our ability to keep invaluable content accessible. Unless timely migration to newer technologies, operating systems and software platforms is assured, we face the risk developing digital Alzheimer's.

UNESCO's expectation from this Conference was to obtain a better definition of our expected role, and our contribution to setting a global digital agenda. The UNESCO/UBC Vancouver Declaration sets out specific recommendations which we will be implementing and incorporating into our digital strategy. Likewise, we expect that our Member States, professional organizations and private sector bodies will also implement the recommendations addressed to them.

Only through collaborative strategic alliances can we overcome the major challenges threatening the preservation of digital information. We believe that the presentations featured in this publication provide the basis for a global commitment to preserving the memory of our world in this digital age.

Jānis Kārklīš
Assistant Director-General
for Communication and Information

Contents

Preface	4
Opening Keynotes	
Kenneth Thibodeau <i>Wrestling with Shape-Shifters: Perspectives on Preserving Memory in the Digital Age</i>	15
Luciana Duranti <i>Trust and Conflicting Rights in the Digital Environment</i>	24
Anne Thurston <i>Digitization and Preservation: Global Opportunities and Cultural Challenges</i>	31
Intellectual Property Infrastructure Initiatives for Digital Heritage	
Heather Christenson and John P. Wilkin <i>Intellectual Property Rights & the HathiTrust Collection</i>	39
Elizabeth Townsend Gard <i>The Durationator® Copyright Experiment</i>	46
Kate Hennessy <i>The Intangible and the Digital: Participatory Media Production and Local Cultural Property Rights Discourse</i>	58
Preservation Infrastructures: Current Models and Potential Alternatives	
Ilaria Pescini and Walter Volpi <i>An Example to Follow: An Infrastructure for Interoperability and Governance in the Tuscan Public System for Digital Preservation</i>	70
Francis G. Mwangi <i>The Road to Providing Access to Kenya's Information Heritage: Digitization project in the Kenya National Archives and Documentation Service (KNADS)</i>	83
Jeremy York <i>A Preservation Infrastructure Built to Last: Preservation, Community, and HathiTrust</i>	92
Hrvoje Stančić, Arian Rajh and Ivor Milošević <i>"Archiving-as-a-Service": Influence of Cloud Computing on the Archival Theory and Practice</i>	108
The CODATA Mission: Preserving Scientific Data for the Future	
Elizabeth Griffin and the CODATA DARTG Team <i>Recovering the Forgettery of the World</i>	127
Patrick C. Caldwell <i>Tide Gauge Data Rescue</i>	134
Stephen Del Greco <i>Environmental Data Through Time: Extending The Climate Record</i>	150
Tracey P. Lauriault and D. R. Fraser Taylor <i>The Map as a Fundamental Source in the Memory of the World</i>	160

Preserving Tradition and Performing Arts in Digital Form

Ravi Katikala, Kurt Madsen and Gilberto Mincaye Nenquimo Enqueri

Life at the Edge of the Internet: Preserving the Digital Heritage of Indigenous Cultures..... 190

Lekoko Kenosi

Digital Madness, Archival Theory and the Endangered Sound Archives of Radio Botswana 206

Jørgen Langdalen

Editing Historical Music in the Age of Digitization..... 212

Lauren Sorensen and Tanisha Jones

Developing and Implementing a Digital Video Repository for Legacy Dance Documentation: Dance Heritage Coalition's Secure Media Network..... 217

Beyond Access: Digitization to Preserve Culture

Fernanda Maria Melo Alves, José António Moreira González and José Manuel Matias

Safeguarding of the Portuguese Language Documentary Heritage: The Lusophone Digital Library..... 229

Benoit Ferland et Tristan Müller

Le réseau francophone numérique 236

John Van Oudenaren

The World Digital Library..... 246

Strategies for Building Digital Repositories

Bronwen Sprout and Sarah Romkey

A Persistent Digital Collections Strategy for UBC Library 257

Neil Grindley

Building the Business Case for Digital Preservation..... 269

Kevin Bradley

Requirements of a Remote Repository 278

Digital Forensics for the Preservation of Digital Heritage

Wayne W. Liu

Accountability for Archival Digital Curation in Preserving the Memory of the World..... 288

Christopher A. Lee and Kam Woods

Automated Redaction of Private and Personal Data in Collections: Toward Responsible Stewardship of Digital Heritage..... 298

Corinne Rogers and Jeremy Leighton John

Shared Perspectives, Common Challenges: A History of Digital Forensics & Ancestral Computing for Digital Heritage..... 314

Giving a Permanent Digital Voice to the Silenced

Terry Reilly

For the Children Taken: The Challenge to Truth Commissions in Building digital collections for research and long-term preservation 338

National Strategies as the Foundation of Togetherness

Andris Vilks and Uldis Zariņš

National Planning as the Key for Successful Implementation of Digitization Strategies..... 348

Ivan Chew & Haliza Jailan

Preserving the Crowdsourced Memories of a Nation: The Singapore Memory Project 354

Ernesto C. Bodè

Digital Preservation Policy of The Chamber of Deputies: Methodology for its development..... 366

Web 2.0 Products as Documentary Digital Heritage: Can We Access and Preserve Them?

Jamie Schleser

Unprotected Memory: User-Generated Content and the Unintentional Archive..... 378

Heather Ryckman

Context 2.0: User Attitudes to the Reliability of Archival Context on the Web 393

Lisa P. Nathan and Elizabeth Shaffer

Preserving Social Media: Opening a Multi-Disciplinary Dialogue..... 410

The Role of Culture in Digitization and Digital Preservation

Fiorella Foscari, Gillian Oliver, Juan Ilerbaig and Kevin Krumrei

Preservation Cultures: Developing a Framework for a Culturally Sensitive Digital Preservation Agenda..... 419

Tukul Sepania Walla Kaiku and Vicky Puipui

Political, Cultural and Professional Challenges for Digitization and Preservation of Government Information in Papua New Guinea: An Overview 431

Xincai Wang and Yunxia Nie

Current Situation, Problems and Prospects of the Digital Preservation of Documentary Heritage in China 439

Open Archival Information System Reference Model: Answer or Inspiration?

Stefano S. Cavaglieri

Digital Archiving Systems Confronted with the OAIS Reference Model..... 451

Saeed Rezaei Sharifabadi, Mansour Tajdaran and Zohreh Rasouli

A Model for Managing Digital Pictures of the National Archives of Iran: Based on the Open Archival Information System Reference Model 457

Collaboration in Digital Preservation or Lack Thereof: What Works

Maria Guercio

Digital Preservation in Europe: Strategic Plans, Research Outputs and Future Implementation. The Weak Role of the Archival Institutions..... 467

Rolf Källman

Models for National Collaboration: Coordination of the Digital Cultural Heritage in Sweden..... 482

Victoria Reich

Building and Preserving Library Digital Collections Through Community Collaboration..... 489

Steve Knight

National Library of New Zealand, Digital Preservation and the Role of UNESCO..... 500

The Economics of Preserving Digital Information

David S. H. Rosenthal, Daniel C. Rosenthal, Ethan L. Miller, Ian F. Adams, Mark W. Storer and Erez Zadok

The Economics of Long-Term Digital Storage 513

Ulla Bøgvad Kejser, Anders Bo Nielsen and Alex Thirifays

Modelling the Costs of Preserving Digital Assets 529

L.M. Udaya Prasad Cabral

Economically Easy Method to Digitize Oversized Documents with Special Reference to Ola Leaf Manuscripts in Sri Lanka 540

Patricia Liebetrau

Preserving Our Heritage: An Independent Advantage 549

Is A New Legal Framework Required for Digital Preservation or Will Policy Do?

Tony Sheppard

Is a New Legal Framework Required for Digital Preservation or Will Policy Do? Building a Legal Framework to Facilitate Long-term Preservation of Digital Heritage: A Canadian Perspective 559

Alicia Barnard

Development of Policies and Requirements for Ingesting and Preserving Digital Records Into a Preservation System: Where to start? 570

Jason R. Baron and Simon J. Attfield

Where Light in Darkness Lies: Preservation, Access and Sensemaking Strategies for the Modern Digital Archive 580

Elaine Goh

Strengthening the Regulatory Framework in a Digital Environment: A Review of Archives Legislation 596

Digital Curation: Convergence of Challenges, Institutions and Knowledge

Sarah Higgins

Digital Curation: The Challenge Driving Convergence across Memory Institutions 607

Jackie R. Esposito

Digital Curation: Building an Environment for Success 624

Patricia Forget

Célébrations institutionnelles : Événement catalyseur propice à l'implantation d'un projet de conservation du patrimoine numérique permettant de réunir les acteurs d'intérêts divergents 636

Jeannette A. Bastian and Ross Harvey

The Convergence of Cultural Heritage: Practical Experiments and Lessons Learned 650

Digitization and Digital Preservation Experiences in a Developing Country Perspective

Elizabeth F. Watson

The Conservation and Preservation of Heritage in the Caribbean: What Challenges Does Digitization Pose? 661

Richard Marcoux, Laurent Richard and Mamadou Kani Konaté <i>Digital Preservation of Demographic Heritage: Population Censuses and Experiences in Mali and the Democratic Republic of the Congo</i>	672
Brandon Oswald <i>Partnership in Paradise: The Importance of Collaboration for Handling Traditional Cultural Expression Material in the Pacific Islands</i>	685
Ensuring That it Won't Happen Again	
Victoria L. Lemieux <i>Financial Records and Their Discontents: Safeguarding the Records of our Financial Systems</i>	700
Myron Groover <i>The White House E-Mail Destruction Scandal of 2007: A Case Study for Digital Heritage</i>	713
Kenneth Thibodeau <i>The Perfect Archival Storm: The Transfer of Electronic Records from the G.W. Bush White House to the National Archives of the United States</i>	724
Trusting Records	
Lorraine Dong <i>The Ethical and Legal Issues of Historical Mental Health Records as Cultural Heritage</i>	735
Marie Demoulin et Sébastien Soye <i>L'authenticité, de l'original papier à la copie numérique : Les enjeux juridiques et archivistiques de la numérisation</i>	745
Web Archiving as Part of Building the Documentary Heritage of Our Time	
Liu Hua, Yang Menghui, Zhao Guojun and Feng Huiling <i>Chinese Web Archiving and Statistical Analysis on Chinese Web Archives</i>	765
Gustavo Urbano Navarro <i>Implications of the Web Semantization on the Development of Digital Heritage</i>	775
Matt Holden <i>Preserving the Web Archive for Future Generations: Practical Experiments with Emulation and Migration Technologies</i>	783
Technology as the Mediator of Heritage and Its Relations with People	
Ian S. King <i>The Turtle At The Bottom: Reflections on Access and Preservation for Information Artefacts</i>	797
Erik Borglund <i>Challenges to Capture the Hybrid Heritage: When Activities Take Place in Both Digital and Non-Digital Environments</i>	814
Limited Resources or Expertise: Case Studies in Addressing the Issue	
Jean Bosco Ntugirimana <i>La problématique de la préservation de la mémoire collective au Burundi à l'ère des NTIC : Étude de cas menée à la Cour supreme</i>	823
Farah Al-Sabah <i>Digitizing A Survivor's Identity: The Past, Present, and Future of the Kuwait National Museum Archives</i>	838

Wayne W. Torborg, Theresa M. Vann and Columba Stewart <i>The Challenges of Manuscript Preservation in the Digital Age</i>	851
Plenary 3 Keynotes	
Dietrich Schüller <i>Challenges for the Preservation of Audiovisual Documents: A General Overview</i>	863
International Perspectives and Cooperation	
Claudia Nicolai, Rachele Oriente and Fernando Serván <i>One Year of Efforts for Digital Preservation at FAO</i>	871
Peter Burnhill, Françoise Pelle, Pierre Godefroy, Fred Guy, Morag Macgregor and Adam Rusbridge <i>Archiving the World's E-Journals: The Keepers Registry as Global Monitor</i>	880
The World Audiovisual Memory: Practical Challenges, Theoretical Solutions?	
Jean Gagnon <i>Treasures That Sleep: Film Archives in the Digital Era</i>	892
Caroline Frick <i>Seeing, Hearing, and Moving Heritage: Issues and Implications for the World's Audiovisual Memory in the Digital Age</i>	896
Edoardo Ceccuti <i>The Digitization of Films and Photos of the Istituto Luce</i>	904
Adam Jansen <i>Challenges and Triumphs: Preserving HD Video at the UBC School of Journalism</i>	909
Mick Newnham, Trevor Carter, Greg Moss and Rod Butler <i>Digital Disaster Recovery for Audiovisual Collections: Testing the Theory</i>	921
Metadata and Formats for Digitization and Digital Preservation	
Joseph T. Tennis <i>Data, Documents, and Memory: A Taxonomy of Sources in Relation to Digital Preservation and Authenticity Metadata</i>	933
Adam Rabinowitz, Maria Esteva and Jessica Trelogan <i>Ensuring a Future for the Past: Long-term Preservation Strategies for Digital Archaeological Data</i>	941
Giovanni Michetti and Paola Manoni <i>It FITS the Cultural Heritage! Formats for Preservation: From Spatial Data to Cultural Resources</i>	955
Lois Enns and Gurp Badesha <i>File Viewers: Examining On-the-Fly File Format Conversion</i>	962
Walter Allasia, Fabrizio Falchi, Francesco Gallo and Carlo Meghini <i>Autonomic Preservation of "Access Copies" of Digital Contents</i>	976
A Methodology Framework to Ensure Preservation	
Anca Claudia Prodan <i>Bias and Balance in the Preservation of Digital Heritage</i>	989

Giovanni Michetti	
<i>Archives Are Not Trees: Hierarchical Representations in Digital Environment.....</i>	1002
Göran Samuelsson	
<i>The New Information Landscape: The Archivist and Architect – Drawing on a Common Map?.....</i>	1011
Shadrack Katuu	
<i>Enterprise Content Management and Digital Curation Applications: Maturity Model Connections.....</i>	1025
Christopher J. Prom	
<i>Facilitating the Aggregation of Dispersed Personal Archives: A Proposed Functional, Technical, and Business Model</i>	1042
Digital Objects as Forensic Evidence	
Carsten Rudolph and Nicolai Kuntze	
<i>Constructing and Evaluating Digital Evidence for Processes</i>	1057
Aaron Alva, Scott David and Barbara Endicott-Popovsky	
<i>Forensic Barriers: Legal Implications of Storing and Processing Information in the Cloud</i>	1064
Michael Losavio, Deborah Keeling and Michael Lemon	
<i>Models in Collaborative and Distributed Digital Investigation: In the World of Ubiquitous Computing and Communication Systems</i>	1079
Fabio Marturana and Simone Tacconi	
<i>Cloud Computing Implications to Digital Forensics: A New Methodology Proposal</i>	1093
Andrew F. Hay and Gilbert L. Peterson	
<i>Acquiring OS X File Handles Through Forensic Memory Analysis</i>	1102
Institutional and Inter-Organizational Initiatives in Digitization	
Anup Kumar Das	
<i>Digitization of Documentary Heritage Collections in Indic Language: Comparative Study of Five Major Digital Library Initiatives in India</i>	1126
Ronald Walker	
<i>Digital Heritage Preservation - Economic Realities and Options.....</i>	1139
S. K. Reilly	
<i>Positioning Libraries in the Digital Preservation Landscape.....</i>	1146
Heidi Rosen, Torsten Johansson, Mikael Andersson and Henrik Johansson	
<i>Experiences from Digidaily: Inter-Agency Mass Digitization of Newspapers in Sweden.....</i>	1153
Preserving Images: What Do We Need to Know?	
Adama Aly Pam	
<i>Chemins de la mémoire : Les archives audiovisuelles au secours de l'identité d'une organisation internationale africaine</i>	1163
Krystyna K. Matusiak and Tamara K. Johnston	
<i>Digitization as a Preservation Strategy: Saving and Sharing the American Geographical Society Library's Historic Nitrate Negative Images.....</i>	1173
Jessica Bushey	
<i>Born Digital Images: Creation to Preservation</i>	1189

Angelina Altobellis

Essential Skills for Digital Preservation: Addressing the Training Needs of Staff in Small Heritage Institutions 1198

Small and Large Scale Digitization: Towards a Shared Conceptual Model

Peter Botticelli, Patricia Montiel-Overall and Ann Clark

Building Sustainable Digital Cultural Heritage Collections: Towards Best Practices for Small-scale Digital Projects..... 1205

Marco de Niet, Titia van der Werf and Vincent Wintermans

Preserving Digital Heritage: The UNESCO Charter and Developments in the Netherlands..... 1219

Paul Conway

Validating Quality in Large-Scale Digitization: Findings on the Distribution of Imaging Error 1233

Lars Björk

Lost in Transit: The Informative Capacity of Digital Reproductions 1252

Preservation of Audiovisual Material

Mike Casey

The Media Preservation Initiative at Indiana University Bloomington..... 1266

George Blood

Video Compression...For Dummies?..... 1273

Pio Pellizzari, Álvaro Hegewich

The Ibero-American Preservation Platform of Sound and Audiovisual Heritage..... 1289

Trusting Data and Documents Online

Junbin Fang, Zoe Lin Jiang, Mengfei He, S.M. Yiu, Lucas C.K. Hui, K.P. Chow and Gang Zhou

Investigating and Analysing the Web-based Contents on Chinese Shanzhai Mobile Phones 1297

Junwei Huang, Yinjie Chen, Zhen Ling, Kyungseok Choo and Xinwen Fu

A Framework of Network Forensics and its Application of Locating Suspects in Wireless Crime Scene Investigation 1310

F.R. Van Staden and H.S. Venter

Implementing Digital Forensic Readiness for Cloud Computing Using Performance Monitoring Tools..... 1329

Yongjie Cai and Ping Ji

Security Monitoring for Wireless Network Forensics (SMoWF)..... 1340

Workshops

Peter Van Garderen, P. Jordan, T. Hooten, C. Mumma and E. McLellan

The Archivemata Project: Meeting Digital Continuity's Technical Challenges..... 1349

Hannes Kulovits, Christoph Becker and Andreas Rauber

Roles and Responsibilities in Digital Preservation Decision Making: Towards Effective Governance..... 1360

Posters and Presentations

Collence Takaingehamo Chisita and Amos Bishi

Challenges and Opportunities of Digitizing and Preserving Cultural Heritage in Zimbabwe 1382

Donna McRostie	
<i>The long and winding road from aspiration to implementation – building an enterprise digitization capability at the University of Melbourne</i>	1384
Asger Svane-Knudsen and Jiří Vnouček	
<i>Retrieving a part of Danish colonial history: From dust to digital copy.....</i>	1386
Mitra Samiee and Saeed Rezaei Sharifabadi	
<i>A Paradigm for the preservation of national digital memory of Iran</i>	1392
Chinyere Otuonye, Tamunoibuomi F. Okajagu, Samuel O. Etatuvie, Emmanuel Orgah, Gift Eyemienbai, Luke Oyovwevotu, Ewoma Borgu, and Janet Ukoha	
<i>Insights on the Digitization of Traditional Medicine Knowledge in Nigeria</i>	1395
Nader Naghshineh and Saeed Nezareh	
<i>Crowd-sourced digital preservation: An Iranian model</i>	1397
Chris Muller	
<i>Data at Risk: The Duty to Find, Rescue, Preserve</i>	1399
Natalia Grincheva	
<i>Digital diplomacy: Providing access to cultural content, engaging audiences on a global scale.....</i>	1401
Rusnah Johare	
<i>Preserving digital research data</i>	1403
Claudia M. Wanderley	
<i>Multilingualism at the University of Campinas.....</i>	1405
Anne Thurston	
<i>Open government and trustworthy records</i>	1407
Jan Marontate, David Murphy, Megan Robertson, Nathan Clarkson and Maggie Chao	
<i>Canada – Aural memories: A case study of soundscape archives</i>	1421
Na Cai, Leye Yao and Liu Liu	
<i>Creating Social Memories of Major Events in China: A Case study of the 5•12 Wenchuan Earthquake Digital Archive</i>	1423
Addendum	
Howard Besser	
<i>Archiving Large Amounts of Individually-Created Digital Content: Lessons from Archiving the Occupy Movement.....</i>	1432
Nadja Wallaszkovits	
<i>Digitisation of Small Sound Collections: Problems and Solutions.....</i>	1440
UNESCO/UBC Vancouver Declaration	
<i>The Memory of the World in the Digital Age: Digitization and Preservation</i>	1452
Sponsors	

Opening Keynotes

Keynote: Wrestling with Shape-Shifters

Perspectives on Preserving Memory in the Digital Age

Kenneth Thibodeau

Abstract

Digital preservation is a difficult challenge due to the polymorphous character of digital information and the environment of ongoing, open-ended and multidimensional change in which it exists. The paper describes both aspects of the challenge and explores how multi-faceted and dynamic approaches to digital preservation in different circumstances can be articulated.

Author

Dr. Kenneth Thibodeau is an internationally recognized expert in electronic records and digital preservation. A senior guest scientist in the Information Technology Laboratory of the National Institute of Standards and Technology of the U.S., he previously directed the Center for Advanced Systems and Technology and the Electronic Records Archives Program at the National Archives and Records Administration in Washington. He also served as the Chief of Records Management at the National Institutes of Health and directed the Department of Defense Records Management Task Force, leading the development of the world's first standard for records management software. Fellow of the Society of American Archivists, Thibodeau won the Emmett Leahy Award and a Lifetime Achievement Award from the Archivist of the United States.

“Words strain, Crack and sometimes break, under the burden, Under the tension, slip, slide, perish, Decay with imprecision, will not stay in place, Will not stay still.”

-- T.S. Eliot¹

The poet T.S. Eliot's passionate incantation of the difficulties of fixing memories in words might be appropriated to describe the difficulties of preserving memory in digital form. Like the raven in North American cultures or the fox of Japanese folklore, digital memory is a shape shifter that takes on very different forms, driven by two distinct causes: first the characteristics of digital information itself and second the environment of change that engulfs digital information objects. There is thus an inherent tension between digital information, which does not stay still, and digital preservation, which quintessentially seeks to keep things in place, without significant change.

1. Polymorphous Information

In contrast to information recorded on stone, clay tablets, paper, or other 'hard copy' media, digital information is polymorphic in several respects. First, digital data is not and cannot be affixed to a physical medium in a durable fashion. Its physical inscription changes every time it moves from computer memory to a storage medium or back, every time it is copied to a different storage medium, and whenever it is transmitted on a network. Digital preservation is not a process of preserving material things, but of

¹ T.S. Eliot, "Burnt Norton," in *Collected Poems, 1909-1962* (New York: Harcourt Brace and Company, 1963), 180.

transporting immaterial bit streams over time, using whatever storage media satisfy preservation needs for however long they are suitable. Most digital storage media are not long lasting, but have to be replaced after some time. The cause of this is the economics of the marketplace, rather than technological or physical constraints on possible storage media. Indeed, several digital storage media have been developed that should last for hundreds or even thousands of years, including microfilm,² metals, gold coated silicone,³ and other formulations;⁴ however, the longevity of the medium is outweighed by the fact that storage devices become obsolete within 5 or 10 years.⁵ Obsolescence in the digital storage domain includes not only equipment that becomes increasingly difficult to maintain and media that wear out, degenerate, and become rare, but also the increasing expense of older storage technologies relative to newer alternatives because of exponential increases in storage density, improvements in data transfer rates, and significant decreases in purchase and operating costs.⁶

A second polymorphic characteristic of digital information is that the boundaries of a digital object can be difficult to determine. For example, web pages often include content that is not visible to the user or that is loaded into the page from external sources each time the page is viewed. External sources include links to other web pages, style sheets, graphic images, Java scripts, data about the person using the page, data elements extracted from databases, and others. Whenever any of these external sources changes, the content of the page changes accordingly, making it difficult to define what is the content of a web page we want to preserve. Moreover, parts of the content of a digital document may be subject to different ownership and control.⁷ In order to preserve a web page, we have to define it as a finite object; that is, we have to apply extrinsic criteria, cutting off at least some external sources of input in order to establish well defined boundaries, but these boundaries are not present in the web page itself.

Furthermore, although many web pages are transitory, many web sites have persisted for decades. The key to this survival is that they are dynamic. They evolve in response to changes in the enabling technologies and also to data about what does and does not work in achieving the purpose each web site is intended to serve. Any attempt to preserve such web sites as static objects loses this essential characteristic of the web site as an evolving entity.

These considerations bring us to a third polymorphic characteristic of digital memories: the relationship between what is stored and what is presented to a human can be both complex and variable. What is presented to a human, as a single object may comprise content drawn from many different data stores, as illustrated in the preceding description of web pages. Databases include rules, invisible to all but administrators that determine what specific data elements different classes of users, or even individual users, can access. Word processing files can contain content that their authors thought they had deleted.

² Heather Brown, John Baker, Walter Cybulski, Andy Fenton, John Glover, Paul Negus, and Jonas Palm. "The role of microfilm in digital preservation," in *DCC Curation Reference Manual*, Digital Curation Center, April 2011. <http://www.dcc.ac.uk/resources/curation-reference-manual/microfilm/>.

³ R. A. Stutz and B. C. Lamartine. "Durable High Density Data Storage," in *Fifth NASA Goddard Space Flight Center Conference on Mass Storage Systems and Technologies*, College Park, Maryland, September 1966, 409-419. http://storageconference.org/1996/papers.html/b2_3.pdf.

⁴ <http://millenniata.com/technology/>.

⁵ Michael C. Peterson, "Solving the Coming Archive Crisis," Storage Networking Industry Association, *SNIA Spring 2007 Technical Tutorials*. <http://www.snia.org/education/tutorials/2007/spring>.

⁶ Chip Walter, "Insights: Kryder's Law," *Scientific American* (August 2005): 32-33. <http://www.scientificamerican.com/article.cfm?id=kryders-law>.

⁷ John Patzakakis and Brent Botta, "Authenticating Internet Web Pages as Evidence: a New Approach," *Next Generation Law and eDISCOVERY Tech Blog*, June 27, 2012. <http://blog.x1discovery.com/>.

Just as something that appears to a human as a single document may be drawn from many data stores, one item of stored digital data may be part of many different objects. For example, a web page may contain links to many other pages, and each of those pages could also be referenced by many others. Conversely, different data can produce identical presentations. A textual document, for example, may be generated from a word processing file, the scanned image of a paper document, or as a report from a database. Thus, in preserving digital memories, we need to distinguish between the *data objects*, which are stored in computer systems, and the *presented objects*, which are derived from the data objects and are presented and at least potentially meaningful to people.

Another polymorphic property of digital memories is that data objects must be processed in order to be used. Moving the data between storage and presentation, or between transmission and presentation can involve changes in semantic, syntactic and apparent form. Even if the data remains intact in storage or transmission, processing for presentation can change or even corrupt the presented object. Furthermore, apart from any question of alteration or corruption, the same digital data can be rendered in different ways; for example, numeric data can be presented in tabular or graphic form. This ability of data objects to take on different shapes may not be merely an incidental possibility. It can be an essential characteristic of the memory we want to preserve. A clear advantage of digital imaging systems in science, medicine, and engineering, for example, is that they allow the data to be presented in a variety of ways. In addition, one of the most prominent aspects of the current digital environment is that much information is intended to be rendered on different types of devices, ranging from various mobile platforms through laptops and desktops to even wall sized and billboard displays. Besides further complicating the distinction between data objects and presented objects, this adaptable display capability contributes to ubiquitous computing, which is changing the role of information and communication technology (ICT) in human affairs.⁸ Thus, preserving data objects is not sufficient for digital memory. We must also maintain the ability to process the data correctly and appropriately.

The polymorphism of digital information means that even the apparently basic issue of what is it that is to be preserved is not a given, but involves choice: should we preserve what was displayed in a given instance or the data, structures, controls, and functionality that enabled the presentation, or both? In order to decide on appropriate choices, we have to consider not only the characteristics of the data objects and presented objects, but also the dynamic context in which digital information exists.

2. An Environment of Change

Digital preservation has a split personality: its object, memory, is from the past but its objective, access, is in the future. This schizophrenia is aggravated by the environment of ongoing, open ended and multidimensional change in which digital information exists.

Ongoing change has two faces, one looking forward, the other backwards. The forward face, technological progress, introduces frequent alterations in both hardware and software that can also include significant innovations or departures. The backward face of ongoing change is obsolescence: older products are no longer supported and become inoperable or unusable, so that, even if we can preserve the data objects, we may not be able process them or to reproduce the presented objects that the data represent. Even in cases where older technology could be maintained, improvements in price/performance

⁸ Adam Greenfield, *Everyware: the Dawning Age of Ubiquitous Computing*, (Berkeley, CA: New Riders, 2006).

of newer products impel us towards replacing it. Obsolescence has been a main focus of attention in digital preservation.⁹ But even if obsolescence were not a factor, changing user expectations about access impel us to alter preservation tactics over time in order to take advantage of technological progress. Moreover, newer technologies may offer better options for preservation, making older preservation solutions themselves obsolete. Technological progress in itself should be anticipated and incorporated in planning for and carrying out digital preservation in order to enable use of the best current technology to preserve, examine, process, and communicate information from the past.

Change in ICT is not only ongoing; it is also open ended, with often surprising developments. The history of ICT since the mid twentieth century is one of repeated transformational changes, where the technology gains new capabilities; new functions are added; new classes of hardware and software are introduced; and even methods of producing and implementing technology change. Software paradigms have shifted from structured to object-oriented, to component-based, and to service-oriented approaches. The emerging paradigm of autonomic computing opens possibilities for additional, radical changes.¹⁰

Technological change has also expanded the varieties of information that ICT can handle. Computation was initially limited to numeric data. Over the last three decades, ICT's scope has grown to include more and more traditional forms of information, such as text, images, audio and motion video. And it has created new forms that cannot exist outside of the digital realm. Additionally, increases in speed and capacity have created new possibilities for processing and communicating information, greatly expanding possibilities for selecting, combining, analysing, and applying different types of information from disparate sources for a variety of purposes. All of this adds considerable diversity and complexity to the challenge of digital preservation.

Moreover, changes in one sphere can snowball into others. Web 2.0 flies in the face of traditional, pre-determined, and systematically controlled user interactions by enabling structure to emerge over time through use of free-form software tools.¹¹ The expansion of mobile computing has spawned the proliferation of "apps," which are substantially changing the end users' acquisition, use, and experience of software, while menacing corporate control of ICT resources.

Moreover, the environment of change is not limited to changes in the technology itself or the types of information it produces. Rather it is multidimensional. First, ICT changes the way we do things. Think, for example, of the differences it has enabled in the interactions between businesses and their customers or between citizens and governments. Second, ICT changes the things we do. For example, geo-positioning technology is enabling precise location tracking of individual vehicles, goods, and people, with substantial impacts on many commercial, governmental and social activities. Third, ICT changes who does what.¹² Prior to the growth of the Internet, for example, advertising was one-way dissemination function, but the possibilities the World Wide Web offers for active customer involvement via social computing has transformed advertising into a multidirectional form of communication in which

⁹ Donald Walters and John Garrett, *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information*, (Washington, D.C.: The Commission on Preservation and Access, 1996); Stewart Brand, "Escaping the Digital Dark Ages," *Library Journal* 124, no. 2 (1999): 46-48.

¹⁰ Richard Murch, *Autonomic Computing*. (Englewood Cliffs, New Jersey: IBM Press, 2004).

¹¹ Andrew McAfee, *What is Web/Enterprise 2.0* <http://www.youtube.com/watch?v=6xKSJfQh89k&feature=related>.

¹² Jay Rosen, "The People Formerly Known as the Audience," in *The Social Media Reader*, ed. Michael Mandiberg (New York: NYU Press, 2012), 13-16.

individual and collective initiative by consumers can have rapid and decisive impact.¹³ The memory of the digital age would be greatly impoverished, and probably falsified, if it does not take into account the additional dimensions of change precipitated by changing technology.

Indubitably, multidimensional and transformative changes will continue in the future, at least as long as ICT continues to change. Inevitably, the challenge of preserving digital memories, and therefore its complexity and difficulty, will evolve apace with changes in ICT and its impact.

3. A Plethora of Choices for Preservation

The polymorphous and metamorphosing characteristics of the challenge of preserving digital memories necessitate multi-faceted, diversified, and dynamic approaches to digital preservation: multi-faceted in order to deal with the polymorphism of digital information; diversified in order to accommodate varying requirements in different social, cultural, and institutional contexts; and dynamic to respond to continuing changes in ICT and its uses and in future user expectations and needs. We can elaborate approaches that take into account the immensity and difficulty of preserving digital memories and that are appropriate to different contexts by addressing three questions: what are you trying to preserve; why are you trying to preserve it; and how much preservation effort is required?

3.1 What are you trying to preserve?

The question of what is to be preserved does not concern selecting things to be preserved, but determining what properties of those things have to survive in order to assert that they have in fact been preserved. Given the polymorphism of digital information, determining the properties that are essential to preserve can be complicated. The possibilities span a spectrum from the preservation of technology to the preservation of information. Between these extremes is the preservation of information artefacts created using the technology. Each alternative responds to different needs and entails different actions.

An obvious case where we would need to preserve information technology would be that of digital artworks that depend on unique technologies. At the other end of the spectrum, for example with statistical data, all we would need to preserve is the information because users could access and use the information with readily available hardware and software. In the middle would be classes like three dimensional models, where we would want to preserve functionality, such as the ability to rotate the model visually, that requires special technology, but where there are alternatives to the original technology used to produce the models.¹⁴

Obviously, determining where digital objects fall on the spectrum of preservation possibilities does not depend solely on the properties of the objects themselves. It is also a function of both the *object class technology* and the available *preservation technologies*. Object class technology includes both the original technology used to produce the data objects and technology currently available for that class of data objects. Preservation technologies are those created to maintain digital memories when the original technology is obsolete and there are no satisfactory alternatives outside of the preservation realm.

¹³ Shuai Yuan, Ahmad Zainal Abidin, Marc Sloan, and Jun Wang, "Internet Advertising: An Interplay among Advertisers, Online Publishers, Ad Exchanges and Web Users," 2 Jul 2012. <http://arxiv.org/pdf/1206.1754v2.pdf>.

¹⁴ Peter Bajcsy, Appraisal of 3D Data Conversions and Visualization Software Packages, January 21, 2009. <http://www.archives.gov/applied-research/ncsa/>.

Illustrating how technology impacts where objects fall in the spectrum, the class of geographic information systems (GIS) is distinguished from mere geographic data by the ability to display data in a cartographic presentation having selected that data from many different types of data stored separately. In principle, as long as we can select data from the separate layers in a GIS and display them in map form, we do not need to preserve the original GIS software. If there were interoperability across GIS formats, GIS would be situated in the middle of the spectrum. However, most GIS depend on proprietary software, which often has features not present in other products in the same object class. If these features were deemed necessary to preserve, maintaining the technology that supports the unique functionality would be necessary, pushing GIS to the technology end of the spectrum.

A given class of digital objects could move across the spectrum, even from one end to the other. In the early days of word processing, for example, visual display technology was too crude to present text in different typefaces. In the 1980s displays were introduced that enabled text to be displayed on screen just as it would be printed on paper, but at that time the only way to preserve that presentation capability was to maintain the specific display technology. Today, however, we do not need any special technology to present text documents from the 1980s with their original formatting.

In sum, then, digital information objects fall at the ‘preserve technology’ end of the spectrum when the only way to ensure continuing access to them is to preserve the original technology or some equivalent or surrogate, such as an emulator. Objects fall at the ‘preserve information’ end of the spectrum when, given their physical survival, they can be accessed using readily available current technology. Objects fall within the ‘preserve information artefacts’ range when they require specialized processing capabilities to render the data objects, but there are alternatives to using the original technology.

In some situations, more than one approach to digital preservation might be appropriate. In the case of interactive digital artworks, where the audience or spectators are involved in real time in the production or performance of the creative work, we face the alternatives of preserving the technology that makes the experience possible or somehow capturing the performance at some particular time and preserving that. In the first alternative, preserving the technology used in performing the work, we would not be preserving memories of specific happenings, but the capabilities that make such happenings possible. In the second, we would not be preserving the digital artwork, but a derivative product that not only does not include any of the technology of the artwork, but also is bereft of precisely what made the art interactive and creative. Basically, this is no different than the alternatives of preserving a written musical score and preserving a recording of a performance of the score.

We also face the option of preserving technology or preserving information in the case of websites where inputs from external sources constantly change what is presented. To enable people in the future to appreciate how users could interact with such a website, we would need to preserve the technology of the website, but that would provide no knowledge of what any users actually saw or might have seen on the website at any time in the past. For that, we would need to preserve snapshots of the website. We cannot decide which alternative is appropriate solely by considering the properties of the information and related technology. For that we have also to address the second question, the purpose of preservation.

3.2 Why are you trying to preserve it?

In addition to the properties of the information objects themselves and the state of related technology, determining the appropriate course of preservation actions depends on the purpose for which digital

information is preserved. We can distinguish two basic reasons for preserving information: either for remembrance or for utilization. We preserve information for remembrance in order to provide the future with opportunities to gain knowledge of the past from materials produced or acquired at the time of which we wish to gain knowledge. We preserve information for utilization in order to enable future use of that information for purposes that are likely to be different from the purposes for which they were created or acquired. While remembrance and utilization are not essentially contrary to one another, they lead us to different preservation actions. Preserving for remembrance would lead us to maintain digital memories as pristine as possible. This goal is best served in many cases by preserving the original information technology. But preserving information embedded in specific technologies creates barriers to exploiting this information in the future. To optimize possibilities for utilizing preserved memories, we would want to reduce or eliminate dependencies on the hardware and software originally used to produce and/or retain the information.

Decisions on what we are preserving and why we want to preserve it are interrelated. This is illustrated in the case of preserving email and other communications on the Internet. In order to work globally, email has to be independent of specific hardware and software as well as relatively impervious to technological change. Thus, there is no need to preserve information technology to preserve email messages. Several initiatives have approached email as an artefact of technology, maintaining the organization of messages within individual users' accounts, because that is the way the technology is implemented.¹⁵ However, if we focus on the value of the information in email as evidence of the conduct of human affairs, the emphasis shifts to the communications between and among individuals and groups of individuals. The threads of communication that evince and often enable important developments in human history are outside of the confines of individual users' accounts and even of the administrative domains of email implementations. To preserve the memory of events ranging from the Arab Spring¹⁶ to the international outpouring of charity on behalf of an elderly school bus monitor who was abused by schoolboys in New York State this spring,¹⁷ we need to preserve the connections among the messages, independently of the artefacts of the enabling technologies.

3.3 How much preservation effort is needed?

What is done to preserve digital memories also depends on how much effort is required. The amount of effort is proportional to the level of resources required to accomplish it. In most cases, resources will be the independent variable. Resource limits may have major impact in determining what is preserved and what preservation actions are carried out. Three sub-factors determine the amount of preservation effort required: quantity, variety, and range.

How much information? In the digital realm, quantity should be measured in both the volume of digital data to be preserved and the number of discrete objects the data comprise, and the probability of substantial growth in both parameters should be a major concern. Between 2006 and 2011, the amount of

¹⁵ The Collaborative Electronic Records Project, <http://siarchives.si.edu/ceerp/>.

¹⁶ Ekaterina Stepanova, "The Role of Information Communication Technologies in the 'Arab Spring' Implications Beyond the Region," George Washington University, PONARS Eurasia Policy Memo No. 159, May 2011. http://www.gwu.edu/~ieresgwu/assets/docs/ponars/pepm_159.pdf.

¹⁷ Rene Lynch, "Bullied school bus monitor calls it quits: She's retiring," Los Angeles Times, July 27, 2012. <http://www.latimes.com/news/nation/nationnow/la-na-nn-bullied-school-bus-monitor-retires-20120727,0,6307218.story>.

digital data produced worldwide doubled every two years, exceeding a trillion gigabytes in 2011, and it is expected to increase fifty times more by 2020. The number of files containing this data has increased even faster and is expected to increase seventy-five times by 2020.¹⁸ These staggering numbers have both global and local implications: there should be significant benefits from international coordination to avoid wasting resources on duplicative efforts and to promote the development of technical capabilities that can be widely implemented; technical solutions developed for particular preservation challenges need to be scalable to accommodate projected growth; and repositories need to anticipate growth in data volumes and numbers of objects that will eclipse everything they have faced until now. The only exceptions would be closed collections, where there will be no further additions.

Given the probability of growth, the quantity of information to be preserved will probably become an increasingly important factor in deciding what gets preserved and how much effort is expended on any set of objects over any period of time. In cases of very valuable information resources, preservers may have to settle for merely ensuring the physical survival of data objects because there will not be resources and perhaps not even any technical possibilities for addressing other requirements, such as overcoming obsolescence or enhancing access.

What variety of digital memories is preserved? Preservation efforts will demand more resources and become more complex as the variety or heterogeneity of the information objects being preserved increases. In general, the greater the variety of objects being preserved, the greater the variety of preservation tactics that will be needed. Homo- and heterogeneity of digital memories are determined at several levels. At the highest level, they relate to the types of information; such as, text, image, audio, motion video, and so on. In addition, in the digital realm, any type of information can be represented in different ways; e.g., text may be encoded as characters or as images of printed documents, and graphic information may be represented by raster or vector data. So, for any type of information, we need to know what data types are used to express it in digital form. Going another level down, a given data type can be encoded in a variety of formats. For example, character encoded text may be in plain text, rich text, Hypertext Markup Language (HTML), eXtensible Markup Language (XML), portable document format (pdf), Microsoft Word, Apple Pages, and other formats.¹⁹

Furthermore, a digital object may be more or less complex. A digital document may consist entirely of textual information all encoded in a single format, but it might comprise several types of information; such as, text, photographs, graphic illustrations, and even audio. A Web page may include static text, data drawn dynamically from a database, images, applets that enable interaction with users, et al.

The variety of formats, and the complexity of objects are also likely to increase the variety of hardware and software necessary to support them. For example, preserving purely numeric data can be very simple; however, if the data are embedded in an object where the specific content is determined in real time from user input, it may be necessary to preserve the database management system used to manage the data, as well as technology needed to reproduce the corresponding presented objects.

What is the range of preservation efforts? Preservation efforts may be directed only at objects individually or extend to preserving relationships among objects. It is not a question of whether objects are related. Basically, all objects subject to a given preservation regime are related simply by virtue of

¹⁸ John Gantz and David Reinsel, *Extracting Value from Chaos*, 2011. http://www.emc.com/digital_universe.

¹⁹ John C. Bennett. "A Framework of Data Types and Formats, and Issues Affecting the Long-term Preservation of Digital Material," *JISC/NPO Studies on the Preservation of Electronic Materials*. British Library and Information Report 50, Version 1.1. 1999. <http://www.ukoln.ac.uk/services/papers/bl/jisc-npo50/bennet.html>.

being held in the same repository or being part of the same collection. Generically, we should distinguish cases where things are simply compiled or grouped together, with no intrinsic ordering or relationships among them, from combinations where there are relationships that must be preserved along with the individual objects. In the case of ‘records’ as defined in archival science; that is, documents produced or acquired and kept in the course of activity, there are essential relationships among records of a given activity; for example, between a letter and the response to it, and between a plan and documents produced in executing the plan and evaluating its success. If these archival links are lost, the possibility of reconstructing the activity on the basis of the evidence provided by the records is diminished.

For purposes of digital preservation, what matters is not the existence of relationships, but whether the relationships require preservation efforts in addition to those required for individual objects. Consider a set of digital maps produced by scanning printed maps. There would certainly be relationships among the maps if they were all pages in a printed atlas, and we would need to ensure that both the relationship of parts to the whole and the sequence of maps in the atlas were preserved; however, these relationships could adequately be preserved in metadata. That would not require any special digital preservation efforts. In contrast, consider the maps that could be produced from a geographic information system. No amount of metadata would be sufficient to preserve a GIS. Many GIS contain such a rich store of data and provide so many options for displaying the data that it would not be possible even to enumerate the set of maps that could be produced using the system. Furthermore, the option of simply preserving each of the various data types or “layers” included in the GIS would not be sufficient because it would not preserve the essential ability to select data elements both within and across layers for composite display in cartographic form. A GIS, which normally consists of cartographic and attribute data, might also be linked to other types of information, such as scientific observations made at specified locations, or historic photographs taken at different times. Preserving such systems requires maintaining the links to such heterogeneous types of information and maintaining the ability to locate them correctly in presented objects.

4. Conclusion

Obviously, digital preservation constitutes an enormous and difficult challenge, one which we must attack less we fail to address important cultural, educational, scientific, social, governmental, and practical needs which depend on, or would benefit from, access to digital memories. The three questions of what, why and how much combine to form a framework for rational discourse on digital preservation, one that embraces the polymorphic character and metamorphosing context of digital information. The framework should guide an integrated and parallel consideration of the three questions because their answers will often be interdependent. This framework should be useful in a variety of contexts, ranging from articulating a theory of digital preservation; to developing a specialized bodies of knowledge and skills, such as digital curation and archival engineering; planning for and managing repositories; developing and implementing strategies for particular sets of information objects to be preserved, and defining the need for preservation technologies, guiding their development, and evaluating the relevance and adequacy of specific preservation techniques.

Keynote: Trust and Conflicting Rights in the Digital Environment

Luciana Duranti

School of Library, Archival and Information Studies, The University of British Columbia, Canada

Abstract

While evolving and emerging digital technologies serve the needs of governments, businesses and individuals to great advantage, the often unintended consequences of their use may be harmful. When WikiLeaks began publishing the largest set of confidential documents ever released, it exposed how endangered are our cherished, yet sometimes conflicting rights—secrecy vs. transparency, privacy vs. access—in the digital world. Moreover, making, storing and accessing records in the highly networked, easily hacked environment of the Internet, is creating liabilities that institutions may not have thought they were assuming. Can the data be trusted? Can the documents from which the data are derived be trusted or even traceable? Are they complete? Are they authentic? Who has access to them? How secure are they? The overview of these and other legal challenges provides a framework for the presentations discussing them.

Author

Luciana Duranti is Chair of Archival Studies at the University of British Columbia, and a Professor of archival theory, diplomatics, and the management of digital records in its master's and doctoral archival programs. She is Director of the Centre for the International Study of Contemporary Records and Archives (CISCRA) and, among many research projects about the issues presented by digital records, InterPARES, Digital Records Forensics, and Records in the Clouds. She is co-Director of "The Law of Evidence in the Digital Environment" Project. Duranti is active nationally and internationally in archival associations and committees, such as the UNESCO International Advisory Committee of the Memory of the World Program; and has been the President of the Society of American Archivists, of which she is a Fellow. She publishes widely on archival theory and diplomatics.

"Whatever matters to human beings, trust is the atmosphere in which it thrives"

*Sissela Bok*¹

On November 28, 2010 WikiLeaks began publishing the largest set of confidential documents ever released, provoking the outrage of governments worldwide, regardless of the many individual voices claiming the morality of such action. Revelation of secret documents is nothing new. What is new is the scale of the phenomenon. Technology has allowed for the uncontrolled growth of databases that can be accessible from any distance. With the amount of data/documents/records created and maintained in digital form, there is a new social awareness of their information potential, and the ease with which they can be disseminated highlights the vulnerability of all parties involved. This fact in itself is at the root of a redistribution of power—the right to know is becoming the core of a new form of democracy that refuses to be held captive to old mechanisms. The WikiLeaks model is destined to spread. TradeLeaks and BrusselLeaks are examples, claiming to reveal frauds in commerce and in the political dealings of European Union members, but in the process, risking damage to the rights of individuals, organizations

¹ Sissela, B. (1978). *Lying*. New York: Pantheon Books.

and governments, as well as to their legitimate operations. These developments are resulting in demands for increased security surrounding digital information, but technology is not the whole answer. *The challenge is providing transparency while protecting the arcana imperii (state secrets).*

In an interesting twist, WikiLeaks entrusted the selection and dissemination of the information to five newspapers for the purpose of avoiding making available data that would hurt military operations or human beings. The old press put its authority at the service of rights to transparency and access by helping certify the reliability and authenticity of the documents, a function of vital importance when their origin is not known and the accuracy of their content may be in doubt. *The challenge is establishing documents' accuracy, reliability and authenticity and maintaining it over time in such a way that it can be proven.*

It is worth noting that Sweden, which in 1766 passed the oldest freedom of information law, is the country that most fiercely condemned the WikiLeaks disclosure. But condemning an action does not prevent its repetition. Iceland recently approved a law that allows for the publication of secret documents. Germany is following suit and so are other countries. Developing new legislation for access requires a profound understanding of the digital environment, of the information generated within it and the various forms it takes, and the way it relates to actions, transactions and facts. *The challenge is to develop legislation and procedures based on an understanding of the way in which digital records serve and protect the rights of the people and of those who govern them.*

In 2009, the Information Commissioner of Canada, in a report entitled *A Dire Diagnosis for Access to Information in Canada*, wrote: "The poor performance shown by institutions is symptomatic of what has become a **major information management crisis** (emphasis original). A crisis that is only exacerbated with the pace of technological developments. Access to information has become hostage to this crisis and is about to become its victim. There is currently no universal and horizontal approach to managing or accessing information within government. Some institutions don't even know exactly what information they are holding."² *The challenge is to develop an infrastructure that ensures a seamless controlled flow of authentic data/documents/records from the creator to the preserver irrespective of changes in technology.*

The right of societies to an enduring documentary heritage became the mission in 1992 of the UNESCO Memory of the World Program. The program inscribes in its registers the records of human achievements as well as those of the darkest moments of human history. It is now grappling with the development of guidelines for the preservation of nominated digital material, which will enable custodians to ensure its continuing authenticity and reliable permanent preservation. *The challenge is to provide guidance to countries, organizations and individuals with different resources and from different cultures, connected by the Internet but divided by their ability to realize its potential while protecting themselves from its risks.*

The challenges described show that several conflicting rights, directly linked to the creation, management and preservation of data, records and archives, are at risk in the digital environment: **the right to transparency and to secrecy, the right to access and to privacy, the right to knowledge and to economic gain, the right to dissemination of one's work and to its integrity, the right to memory and to right to be forgotten, the right to the endurance of one's heritage and the right to oblivion.**

² Office of the Information Commissioner of Canada. (2009). *A Dire Diagnosis for Access to Information in Canada*. Online: http://www.oic-ci.gc.ca/eng/med-roo-sal-med_spe-dis_2009_4.aspx.

How can we protect these conflicting rights? Whom and what can we trust with the care of the digital objects that embody them, attest to them, support them, result from them, are the object of their exercise, or disseminate them so that they can be nurtured, respected, guaranteed, and regarded as certain and clear? The certainty of people's rights as objectified in the world's documentary residue is one of the pillars of every democratic society. As Baldassare Bonifacio wrote in 1630:

There is nothing more necessary for clearing up and illustrating obscure matters...for conserving patrimonies and thrones, all things public and private, than a well constituted store of volumes and documents and records—as much better than navy yards, as much more efficacious than munitions factories, as it is finer to win by reason rather than by violence, by right than by wrong.³

And as Sir Hilary Jenkinson re-stated three centuries later, documents are “the material evidence of the historical case.”⁴ Again, whom or what do we entrust with this invaluable store of rights?

Traditionally, trust in documents has been based on trust in those who hold them in custody. The grounds for it are: *reputation*, which results from an evaluation of the custodians' past actions and conduct; *performance*, which is the relationship between the custodian's present actions and the conduct required to fulfil his or her current responsibilities; *competence*, which consists of having the knowledge, skills, talents, and traits required to be able to perform a task to any given standard; and *confidence*, which is an assurance of expectation of action and conduct.⁵ With respect to the digital material produced by contemporary society in both the public and private sphere, do we still have confidence in the competence, performance and reputation of those who have it in their custody? If we do, should we? Is the legal framework in which they operate strong enough to ensure that our trust is well placed?

In contemporary practice, individuals and organizations are increasingly saving and accessing records in the highly networked, easily hacked environment of the Internet, where current policies, practices and infrastructure prohibit us from being able to assess our trust in records relying on the kind of understanding we used in the past. How do we know that those who hold digital records about us make the right decisions about keeping them safe, and accessible only to those who have a right to see them, using them for good and in a transparent way, disposing of them when required, and selecting reliable Internet providers for storing and managing them? Who has established the rules according to which they operate, and in the context of what values and purpose?

The interconnectedness of the Internet is forcing us into one community without the benefit of gradually getting to know one another. As the United States developed the Internet, its social, political, and economic views are reflected in its management, thereby rankling other countries. A recent example highlights the risks of using a consumer file-sharing service for business purposes when it is not clear what legal framework controls it. U.S. federal prosecutors blocked access to the file-sharing site Megaupload.com on charges that the site violated piracy laws, and New Zealand police arrested Megaupload's founder based on the U.S. accusations. As a consequence the data of at least 50 million

³ Born, L. (1941). “Baldassare Bonifacio and His Essay *De Archivis*,” *The American Archivist* IV, 4: 233-234.

⁴ Jenkinson, H. (1980). “The English Archivist: A New Profession,” in *The Selected Writings of Sir Hilary Jenkinson* (Gloucester), pp. 246–47.

⁵ Borland, J. (2009). “Trusting archivists.” *Archivi & Computer*, XIX(1):96–109; Duranti, L. and Rogers, C. (2011). “Educating for Trust,” *Archival Science*, Volume 11, Issue 3, pp. 373-390. Online: SpringerLink doi:10.1007/s10502-011-9152-3. See <http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s10502-011-9152-3>; Sztompka, P. (1999). *Trust*. Cambridge University Press, Cambridge.

Megaupload users ran the risk of being erased.⁶ Convinced that existing laws could not deal with growing piracy concerns, the U.S. Congress introduced the *Stop Online Piracy Act (SOPA)*, which resulted in protests across the Internet that persuaded Congress to reject the bill. In the meanwhile, the European Union proposed a “right to be forgotten” directive, which would have required every member state to issue legislation protecting online intellectual rights and privacy.⁷ Thankfully, this initiative also, was unsuccessful, but it shows how unclear is what is best to do. Google established a blanket privacy policy for all materials on its cloud,⁸ while Twitter chose to go the opposite way and to adopt the policy of the country of origin of the record.⁹ Indeed, the Internet has forced us into one community, but one community in desperate need of a shared legal framework that promotes consistency and balance in terms of policies and practices regarding the handling of digital objects, especially when they reside with Internet services and social media providers.

In fact, regardless of several public cases of dramatic documentary incidents, people in general trust all kinds of organizations, like banks and phone companies, to keep and maintain their data/records/archives on their behalf. In effect they have shifted their trust from the central records repository in their home or office to distributed archives online, the stewardship for which is entrusted to others. Where their records actually reside, how well they are being managed, how long they will be available to them... they have no idea! Many organizations are recognizing this shift and becoming concerned about a liability they may not have thought they were assuming, especially as more and more clients abandon their own recordkeeping, and place greater reliance and trust on the recordkeeping abilities of the organizations with which they interact.

In additions, commercial organizations like telecommunications services, distributors, and the like, are amassing huge volumes of data that they use to provide a host of services, many of which focus on marketing and securing competitive advantage. This is the evolving world of big data’, the exploitation of seemingly innocuous records, like call centre records, purchase orders, etc., to produce data that can be re-manipulated to serve a host of purposes, also called ‘data mining.’ Big data is introducing a view of our documentary output that flips our traditional view on its head: certain records can grow in value if it is recognized that their accumulation through time will enable the production of data that themselves will grow in value as their potential to support organizational priorities—especially strategic priorities—is realized. However, big data also fosters a range of democratic objectives, from promoting government transparency to supporting research to contributing to public-private sector goals and priorities. Thus, legislation, regulations, policies are needed to control these activities so that their benefits can be ripped and their risks contained.

The issues for data and records coincide. Can the data be trusted? Can the records from which the data are derived be trusted? Are they complete? Are they authentic? How were they generated, by whom and under what conditions? Is there sufficient contextual information to enable them to be understood? These are questions faced by quite a number of organizations that are beginning to act on the realization

⁶ Maes, J. (2012). “SOPA, PIPA, Megaupload.com, and the United States Government,” Washington Times (3 February). Online: <http://communities.washingtontimes.com/neighborhood/political-potpourri/2012/feb/3/sopa-pipa-megauploadcom-and-united-states-governme/>.

⁷ <http://www.bbc.co.uk/news/technology-16677370>

⁸ Google (2012). *Preview. Privacy Policy 1 March 2012*. Online: <http://www.google.com/intl/en/policies/privacy/preview/>.

⁹ Twitter Blog (2012). “Tweets Still Must Flow.” Online: <http://blog.twitter.com/2012/01/tweets-still-must-flow.html>.

that their data holdings, and the records generating that data, are digital assets that need to be managed effectively if they are to be trusted by those making decisions and by clients, customers, citizens, etc. One of the catch words in this arena is ‘traceability’, that is, the ability of an organization to trace back from the data it is using for decision-making, service delivery, etc., to the source records from which the data are derived. The issue of traceability of data to trusted records is becoming huge and constitutes the foundation of trust in data.

Different but equally significant issues are generated by the fact that individuals and organizations, large and small, are drawn increasingly by the lure of cloud computing for the many benefits it offers. Scalable, agile, efficient, on-demand computing resources mean that email, photos, documents, records, and archival fonds can be easily stored and shared through a seemingly endless number of hosted web applications, and that sophisticated software, platforms, and infrastructure are available to the budget-conscious and the technology-resource limited. Cloud architectures offer on-demand access to services across a network of standard internet-accessible devices—mobile phones, tablets, laptops—and a vast array of other equipment, such as game consoles, MP3 players, and e-business technologies. Resources are shared among users, and resource use is monitored and invoiced based on usage for service. We choose—and increasingly rely on—cloud services for communication, backup and storage, collaboration, distribution, recordkeeping and preservation. But for every benefit there is a corresponding risk that may or may not be recognized.

The model of cloud computing is reminiscent of the mainframe environment of the 1960s, except that in this case we are not putting our trust in the proprietary and highly controlled environment of the company mainframe, but in global service providers, whose agendas and priorities as they build out their infrastructures are very different from our own. The trust relationship demands careful analysis and consideration and it is important to highlight specific challenges to entrusting data, records and archives to the cloud. Key issues of ownership, jurisdiction, and privacy have yet to be resolved. Longer term concerns around responsibility for maintenance, access, and preservation, all of which correspond to issues of trust, are looming on the horizon. The following list identifies some legal concerns but is by no means exhaustive:

- The servers in which data and records are stored may be, but likely are not, in the same country or jurisdiction in which they were created. In the event of litigation or other dispute, in what jurisdiction will they be governed?
- Do you even know with which provider your material is stored? As the cloud storage market continues to grow, this becomes increasingly unclear. New storage providers are appearing who aggregate unused storage from third parties. The entrance of a peer-to-peer model for storage adds further complexity to teasing out the tangled web of provenance, custody, control, and legal responsibility.
- Will trade secrets or legal privilege, if entrusted to cloud storage, continue to exist after they have been shared with a third party?
- How will cloud service providers protect content from data breaches? There is a school of thought that says you should be concerned not about *if* a data breach occurs, but *when* it occurs. How will your cloud service provider handle a breach? Will your provider even admit to a breach?

- Audit is usually not allowed by cloud providers: How do you prove authenticity and an unbroken chain of custody?
- What happens to content if a cloud service provider goes offline, due to bankruptcy or criminal investigation, or if the server containing your material is sequestered for an investigation?
- How do you transfer the material to the designated cultural institution with independent evidence of chain of custody?

With these questions in mind, returning to the concept of trust and the protection of rights embodied in and exercised through data, documents, records, archives, if trust rests on our confidence in the reputation, performance, competence, and confidence of the custodian of our digital material, we must ask hard questions of those to whom we entrust our data, records and archives. International research projects into the nature of digital records have developed guidelines and solutions to managing authenticity, accuracy and reliability in digital records systems, but solutions are often out of reach financially for many individuals, organizations and countries driven by the bottom line. National and international standards of records and information management provide guidance but adherence is not legally required in most sectors. Cloud computing offers to ease the financial burden of many aspects of records management and archival storage, but—as we have seen—raises a host of new and troubling questions that must be answered if we are to be able to trust and maintain access to our material. Technology will not stand still to wait for our legal and regulatory system to catch up. Even if it did, domestic legislation, as controlled as it is by the higher law of each country, and, in common law countries, also by case law, would often be conflicting with that of other jurisdictions, and, looking at the world map, we would be seeing a tower of Babel of legislations that create more problems than they can solve.

What we need is an internationally agreed upon legal framework that will support the development of integrated and consistent local, national and international networks of policies, procedures, regulations, standards and legislation concerning digital records, to ensure public trust grounded on evidence of good governance. Such legal framework needs to anticipate problems in maintaining any trust in digital data/records/archives which are now under the control of entities suffering a waning level of confidence from the public, including legal, law enforcement, financial, medical, broadcasting, and governmental organizations and professionals, especially in light of the noted exponential growth of and reliance on Internet services. This could be done by means of “model legislation” that can be adapted to each national and cultural context. This model legislation would allow for a harmonization of provisions related to the proper control of our digital heritage from the moment of creation throughout its life-cycle so that it will be produced and maintained in an accurate and reliable way and its authenticity will be protected from the very beginning. A model legislation needs to be detailed enough to contain exemplary norms about specific issues presented by digital material, but general enough to be independent of technological changes, focusing on concepts rather than processes, principles rather than activities.

The only body which can take up the responsibility of writing this non-legally binding model law for the protection of the rights embodied in and exercised through digital documents, and which has the authority and the recognition for doing so is UNESCO. UNESCO can issue a model law as a recommendation, that is, as an instrument in which “the General Conference formulates principles and norms for the international regulation of any particular question, and which invites Member States to take whatever legislative or other steps may be required in conformity with the constitutional practice of each State and the nature of the question under consideration to apply the principles and norms aforesaid within their respective territories” (Article 1 (b)).” The UNESCO *Charter on the Preservation of Digital*

Heritage, whose purpose is to guide the member states in overcoming the challenges of digital preservation, as revised on the basis of the recommendations coming out of this conference, is the ideal contextual document for a model law. As its natural complement, a model law would guide the legislative bodies of the member states in ensuring the proper implementation of the Charter's general guidance through the issuing of domestic legislation. This conference is the time to start thinking what a model law should include. If not now, when?

Keynote: Digitization and Preservation

Global Opportunities and Cultural Challenges

Anne Thurston

International Records Management Trust

Abstract

As countries worldwide enhance their ability to operate in the digital environment, our societies have the greatest opportunity the world has ever known for preserving and sharing information and empowering citizens through access to information. There is a rapidly growing international awareness of the potential benefits but less recognition of what must be in place to realize them. The challenge is to articulate clearly the essential significance of authentic, trustworthy records as the foundation for transparency and accountability along with the risks of failing to address digital records management and preservation; and to share knowledge of good practice across the information profession as rapidly as possible.

Author

Anne Thurston has pioneered approaches to sharing solutions for managing public sector records with developing nations. Between 1970 and 1980 she lived in Kenya where she conducted research before joining the staff of the Kenya National Archives. In 1980 she became a Lecturer, later a Reader in International Records Studies at the School of Library, Archive and Information Studies, University College London. She established the International Records Management Trust in 1989 and continues to be its Director. In 1996 she left University College London to concentrate fully on the work of the Trust. In the 1990s, recognizing the impact of the rapid changes in the use of information technology on the management of public sector records, she structured the Trust to address the impact of those changes. Dr. Thurston was a member of the UK Lord Chancellor's Advisory Council on Public Records from 1994 to 2000. She was awarded an OBE for services to public administration in Africa in 2000 and a Lifetime Achievement Award by the Records Management Society of the UK in 2007. She was awarded the Emmett Leahy Award for Outstanding Contributions to the Information and Records Management Profession in 2007.

1. Introduction

UNESCO and UBC have achieved something extraordinary in convening this conference. Our societies have the greatest opportunity the world has ever known for preserving and sharing information and empowering citizens through access to information. It is the right time, and you are the right people to take this issue forward. As UNESCO notes on the conference website, digital information not only has value as a cultural product and a source of knowledge, but it is essential to sustainable national development as, increasingly, personal, governmental and commercial information is created in digital form only. Digital and digitized records are rapidly becoming the basis for citizens to exercise and protect their rights.

My work over the last 40 years has given me a context for examining these issues. I lived in Kenya for 10 years in the 1980s and joined the staff of the National Archives of Kenya. Later, I taught at University College London and conducted extensive research on the management of government records

in 32 countries worldwide. On this basis, I set up the International Records Management Trust,¹ which I have run for over two decades. The IRMT has worked in virtually all regions of the world but mostly in Africa, where we are committed to providing support to densely populated low-income countries, which we believe can benefit greatly from strengthened records management. Our three-pronged approach, which involves on-going research, consultancy, and development/ distribution of free educational material and assessment tools, has given us an overview of the challenges countries are facing worldwide.

The global wealth of expertise gathered at this conference is extraordinary. At the same time, the lack of high-level support for digital records management and preservation is a stark reality, with a massive widening gap between what is needed and what is being achieved. If the digital memory of the world, including that portion of the memory resident in government records, is to survive, we need to take a careful look at how we position ourselves strategically to bring about change.

UNESCO has challenged us to launch specific initiatives related to digital preservation, to foster access to documentary heritage through digitization, to identify legal frameworks that will facilitate long-term digital preservation, to agree on our approach to exchange standards, and to define the respective roles of professionals, academics, industry and governments in relation to these issues. These are excellent and necessary objectives, but to achieve them, we need to understand why digital access and preservation receive so little recognition and are so poorly funded in most parts of the world. How can we reverse this position? What are the priority development areas to which we can contribute? I believe that in addition to technical competence, we have to position ourselves to make our contribution in the context of the growing global emphasis on citizens' right to transparent and accountable government. In today's climate of limited funding, competing priorities and budget cuts, we have to demonstrate relevance to global needs.

2. Barriers and Challenges for Digital Preservation

In this presentation, I am focusing on government records and data, not because I am unaware of the importance of non-governmental cultural assets or of personal and commercial records, but because few societies invest in cultural preservation for its own sake: there must be practical and demonstrable benefits to the society. Preserving digital records as cultural assets and managing them as organisational assets must begin by addressing the weaknesses in the management of current records. Unless we are positioned close to the point at which the records are still being actively managed in the creating agencies, the records are unlikely to be available, understandable and useable through time. Ultimately, leadership and funding for preservation and access will come largely through governments and the donors and lenders that support them, but a number of crosscutting issues will have to be resolved.

2.1 Issue One: Digital Preservation Is Not a Development Priority

This is, I believe, the single most important issue that we must address. International donor and lender support is often a crucial factor in influencing national development priorities and an essential source of funding, but at present, digital preservation is not even on the radar of the global development community. Unless this changes, the money to pay for digital preservation and the structures needed to support it will not be available. Too frequently, digitization and business process automation are

¹ See www.irmt.org.

concerned only with the present, with little attention to the future. We need to help governments, donors and lenders understand that digitization and preservation are a fundamental part of the context of development.

2.2 Issue Two: Lack of Awareness

For the most part, development planners and government stakeholders are not yet aware of preservation and access issues, of the cost of the failure to address these issues or of what is needed to do so. They still tend to believe that technology will resolve the problems. The information profession has not yet spoken powerfully into the development process. We have not made it clear why and how our profession is such a crucial factor for social and economic development. Introducing Information and Communications Technology (ICT) is a major development goal in countries across the world. Government officials and development planners tend to assume that digital information will survive without intervention. They focus on the dramatic benefits of digital systems without considering the integrity of the digital information that these systems generate. Millions of dollars are invested in information systems, but records are not being captured in a form that will be intelligible, unalterable and usable over time. Government agencies have hardly begun to take responsibility for the records they generate.

2.3 Issue Three: Gaps in the Institutional/ Regulatory Framework

In many countries the combination of laws and policies, standards and practices, enabling technologies and qualified staff needed to ensure that digital records remain accessible and trustworthy are not in place. Legislation is often out of date or inappropriate, and conflicting laws tend to split responsibilities for government records or result in an absence of responsibility, so that no one is accountable. In many cases, international standards have not been introduced, and often planners don't know that they exist. There tends to be an absence of national information policies, and where policies exist, they tend to relate to paper records. Trusted digital repositories rarely exist, and digital records are often stored on various recording media in computer rooms or in rooms with poor environmental controls. In many cases, basic procedures and management controls for digital records management and preservation have not been developed and implemented, and there has been little planning for continued accessibility in the changing ICT environment.

2.4 Issue Four: There Is a Lack of Practical Capacity to Manage and Preserve Digital Records

Relatively few records professionals worldwide have had in-depth training and experience in managing and preserving digital records. University education programmes are beginning to address the issues, but the lack of practical expertise nationally tends to mean that education and training remain theoretical. Newly qualified professionals flounder when faced with the enormous challenges of turning theoretical learning into practical solutions.

2.5 Issue Five: Digitization Initiatives Can Fail Due to a Lack of Preparation and Standards

Digitization is a priority issue for many governments and international organisations as a quick means of making records accessible and ending dependence on paper records. Digitization projects fail or achieve limited results where the original paper records are poorly organised and international standards for

digitization are not recognized. Too often, the management and quality control framework needed to ensure that the digitized records meet requirements for legal admissibility, reliability and authenticity, and long-term preservation have not been developed. Requirements for image resolution, metadata fields, standardised indexes and classification structures, and retention and disposition schedules are often unrecognized. In some cases, agencies assume that hard copy source records can be destroyed as soon as digital copies are created; if the scanned image is poor or there are problems with accessing it, the agency and civil society are at risk. Where the organisation responsible for the digitized records does not have a digital repository and a digital preservation strategy in place, the digitized records are unlikely to survive in the long-term.

3. Consequences of the Current Situation

Meaningful citizen engagement in the digital environment requires on-going access to trustworthy, reliable and accurate records and datasets. If they are not professionally managed in secure technology-neutral facilities and supported by complete metadata, they are unlikely to be available to support citizens' needs. Unmanaged information can be manipulated, deleted, fragmented or lost, and records can become unreliable. Citizens cannot prove unequal or unjust treatment, delivery of justice is impaired, and human rights cannot be protected. Access to Information (ATI) requests cannot be met promptly or accurately. Data drawn from inaccurate and incomplete records can lead to skewed findings and statistics. Governments' ability to make decisions and achieve strategic priorities for economic and social development is seriously undermined. Misuse of information, cover up of fraud, misguided policy recommendations and misused funding all contribute to poor governance. Information technology projects are often not sustainable or do not reach their intended goals because records issues are not addressed as part of the planning and because implementation process and because data is not kept accurate and up to date.

4. Open Government: An Opportunity for the Records Profession

Over the last several decades, governments and donors have tried to improve the quality of governance through strategies and programmes for addressing poverty reduction, structural adjustment, democratisation, service and programme improvement, political regime stability, evidence-based governance, electronic government and anti-corruption. Records are essential for all these objectives, but the development community has not recognised their crucial significance. The Open Government Partnership (OGP),² launched in September 2011, offers a significant new opportunity for our profession.

The Partnership is based on the idea that governments exist for the benefit of the people; the people should have access to information about what their governments are doing so that they can hold them accountable and get the greatest possible advantage from government information. The Open Government Declaration, signed in September 2011, recognised that people around the world are looking for ways to make their governments more transparent, responsive, accountable and effective. This remarkable initiative, involving leaders at the highest political level, endorsed the principles of the Universal Declaration of Human Rights, the UN Convention against corruption and other instruments related to

² See <http://www.opengovpartnership.org/>.

human rights and good governance. To date, 57 countries have joined the initiative, 45 of which have delivered commitments and 12 of which are developing commitments. Moreover, Open Government is being discussed in many countries that have not applied to join the Partnership.

The transparency and accountability field is now one of the fastest growing public movements of recent years. It brings together a wide range of organisations and projects aimed at promoting greater openness on the part of governments, companies and other institutions so that the public can hold them to account. It represents a unique opportunity for the records profession because well-managed records, as evidence of government policies, activities and transactions, are the cornerstone of openness. I do not believe that the movement can truly succeed without our involvement.

At present, however, records barely feature on the Open Government agenda. Instead, Open Data has an increasingly high profile because it is an immediate way of making government information available. Open Data involves opening non-sensitive datasets as a way of promoting transparency, accountability and economic development. The US and UK led this movement initially, but Open Data is rapidly gathering momentum worldwide. In July 2011, Kenya became the first African country to launch an Open Data portal, releasing over 160 datasets including budget and expenditure data, as well as information on healthcare and school facilities. This has caused great excitement in the development community.

For all the excitement about the potential of Open Data, the fact remains that if governments are to prove accountable and achieve their economic and social objectives, and if citizens are to engage meaningfully with their governments in the digital environment, on-going access to trustworthy, reliable and accurate records and datasets is essential; data is only meaningful if it can be traced back to the records from which it is derived. There is a strong relationship between records management and government accountability, decision-making, service delivery and ability to achieve strategic priorities. Citizens and investors need to know they can trust the information that governments provide. When they make requests under Access to Information legislation, they have the right to expect that the information will be provided promptly and will be accurate and authentic. When datasets are released through Open Government portals, they have the right to expect that the data can be trusted. Only then can Open Data and Access to Information become true means of ensuring government transparency and openness.

We have an opportunity to raise the profile of records management significantly, with governments and with donors and lenders, if we can make the case clearly that records are the basis for successful openness and bring this issue onto the Open Government agenda. Getting records management into OGP country action plans that members are required to develop is a valuable opportunity to raise the profile of records management services, argue for greater resources and make a significant contribution to national and international development. Beyond that, we need to work with international partners to get records management and preservation onto the OGP agenda.

5. Bringing Records Issues Onto the OGP Agenda

The IRMT, working with the International Council on Archives (ICA) and the Transparency and Accountability Initiative³ in London, has begun this process. The ICA, as the international NGO dedicated to the effective management and preservation of records and archives, recognises the

³ See <http://www.transparency-initiative.org/>.

opportunity that Open Government offers for the records community. The Transparency and Accountability Initiative, which brings together philanthropic foundations, official aid agencies and civil society networks to promote innovation in transparency and accountability across many fields of international development, has accepted that records are fundamental to openness. This month, together we launched an initiative on Open Government and Trustworthy Records involving an international stakeholder assessment of a narrative and an assessment tool aimed at helping governments set direction for records management and preservation as elements of their OGP action plans.

The Assessment looks at the institutional/ regulatory framework and capacity in place at three levels of achievement. In the *initial steps*, the goal is to ensure that evidence of government decisions, actions and transactions is created, captured and managed in fixed and accessible form as reliable and authentic records, to underpin transparency and accountability. In the *more substantial steps*, the goal is to ensure that records management requirements are addressed in relation to Open Data/ Access to Information requirements and ICT/ e-Government initiatives, and that these requirements are integrated in the design of government business systems. The goal for the *most ambitious steps* is to ensure that proactive disclosure of records and the information and data derived from records is embedded in government processes and cultures, thereby promoting engagement between governments and citizens.

6. An Example of What Can Be Achieved: Norway⁴

Norway's initiatives in records management and digital preservation offer insight into what is achievable. The Norwegian model requirements for electronic records management systems were first introduced in 1984, and over almost 30 years, ongoing developments in the complexity of platforms, system portfolios and functionalities have been fused into a common standard for a wide range of applications.

The model requirements emphasise the direct relationship between quality at the point of creation and the ability to reproduce the records according to their original structure to show relationships and original context in a way that upholds their authenticity, accuracy and context. This is achieved by ensuring that everything is identified, labelled and coherently structured within the model requirements. Legislation narrows down the range of allowed file formats for transmission, and vendors may not sell electronic document and records management systems in the public sector without proving that their solutions comply with the model requirements.

One of the most striking aspects of the Norwegian approach is the awareness that preservation itself is not enough. From a historical perspective, access to archives does not have to happen quickly. However, if the material is needed because it contains important legal, financial, government based or rights-oriented information, it is important to reduce the time between creation and public access by users, including lawyers, courts, public bodies, researchers and individuals. The aim is to meet increasing citizen expectations for rapid, almost real time access to information and to make digital repositories a natural part of the digital environment.

⁴ This analysis is based on extensive communication with Olav Hagen Sataaslaatten, Assistant Director General of the National Archives of Norway. See Olav Hagen Sataaslaatten, "Does our ability to preserve and create future access to data depend ultimately on the quality of model requirements in its creational phase?" (poster to be presented at iPres2012, University of Toronto, October 1-5 October, 2012). See also descriptions of Noark 5, arkivverket.no/arkivverket/Offentlig-forvaltning/Noark/Noark-5/English-version, and of Norway's Electronic Public Records System, www.epsa-projects.eu/index.php?title=Electronic_public_records.

The Ministry of Government Administration, Reform and Church Affairs offers an Electronic Public Records System as part of the Norwegian Government's commitment to transparency and democracy within the public sector. Based on the Freedom of Information Act and related regulations, it aims to make the public sector more open to citizens. Central government agencies use this tool daily to publicise metadata about the records they create online so that users can identify documents relevant to their interests and submit requests to see them. The documents are provided within a matter of days. Within a few years from their creation, the National Archives of Norway receives these records from the ministries through a digital repository structure that meets TRAC and OAIS requirements and reduces the need for manual operations through standardised models for digital preservation. In 2011 when the National Archives developed a 'smartphone app' to enable public access to records and data direct from its repositories, the range of possibilities for information retrieval advanced to a new level.

7. Conclusion

At present, the lack of capacity, appropriate institutional/ regulatory frameworks and funding are substantial impediments to digital records management and preservation. So long as these issues have low priority, the situation will grow more critical as the volume of records in digital form increases substantially in coming years and the records from previous technologies age. Marginal increases in funding will not be enough to reverse the situation. As we examine initiatives for strengthening long-term digital preservation and consider the roles that different stakeholders need to play in this process, it is also crucial to position the information profession in relation to the global development context, particularly the Open Government movement.

We have a unique and essential contribution to make to international development, because we have the means of ensuring that digital information is protected and preserved in a trustworthy, authentic form. We have made tremendous progress in developing the concepts and tools needed to manage digital records, from strategies and standards, to laws, practices and technologies, and we have a powerful global network of information specialists. However, in order to make our rightful contribution, the profession will have to reposition itself, to move forward in new directions, to find new ways of sharing good practice and collaborating with new stakeholders. Government archives, for instance, will need to go beyond guidance and regulations to assume leadership and oversight. They will need to work actively, not only within the professional information community, including libraries and museums, but also with government stakeholders, including those responsible for Access to Information, Open Data, Electronic Government and audit.

Most importantly, the profession will need to focus on becoming relevant to citizens' needs. If we fail to do so, there will be significant losses for government accountability, economic opportunity, citizens' rights and the preservation of knowledge. This may begin with international bodies such as UNESCO and the ICA reaching out to new partners, with universities considering new joint programmes, with individual professionals opening new discussions across government agencies. The bottom line is that if we want to serve the world's citizens by protecting and preserving digital information, if we want the funding situation to change, we must change.

Intellectual Property Infrastructure Initiatives for Digital Heritage

Intellectual Property Rights & the HathiTrust Collection

Heather Christenson¹ and John P. Wilkin²

¹California Digital Library, University of California, USA, heather.christenson@ucop.edu; ²HathiTrust, USA

Abstract

Research libraries founded HathiTrust in 2008. This digital preservation and access collaboration of over 60 research libraries in the United States, Canada, and Europe utilizes a shared infrastructure to preserve digital copies of now over 10 million volumes digitized from print. HathiTrust's mission is "to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge." This paper introduces the goals of HathiTrust, describes the scale and scope of the HathiTrust collection and its significance, and discusses how the organization is providing services related to the digital collection, in light of changing conditions for maintenance of the participating libraries' print collections, and, in particular, in the context of the current environment for intellectual property rights.

Authors

Heather Christenson is Mass Digitization Project Manager at the California Digital Library (CDL), and Project Manager for the University of California's HathiTrust collaborative work. Ms. Christenson is responsible for coordinating operations across the UC Libraries' large scale digitization projects, and plays an integral role in planning for services surrounding digitized content. Prior to joining CDL, Ms. Christenson was involved in the development of commercial web search tools, and was a news librarian, law librarian, and cataloger. She received her M.L.I.S. from the University of California, Berkeley.

John P. Wilkin is the Associate University Librarian for Publishing and Technology at the University of Michigan, and is the Executive Director of HathiTrust. Mr. Wilkin provides operational and strategic leadership for HathiTrust, a growing international partnership committed to using digital collections to help libraries improve access and preservation strategies. In his role at the University of Michigan Library, Mr. Wilkin also leads publishing efforts, including the Library's pioneering digital publishing operations and the University of Michigan Press, and the Library's technology operations.

1. Introduction

Research libraries in the United States have been digitizing their materials for almost two decades, both individually and via collaborative projects. Digitization is expensive, and in the absence of an official national library program or long-term national funding, the libraries have accomplished the task of converting books to digital form in a number of ways: through partnerships such as collaboration with Google, grant funding, and self-funding. Unlike commercial enterprises, however, research libraries place a great deal of value on digital preservation, and in the provision of digital content for scholarly uses into the future.

Although the conversion of library materials from print to digital form has happened at a brisk pace, the law has been slow to evolve in terms of considering use of mass digitized library collections. Research libraries view as an imperative their traditional role as stewards of the record of human knowledge, regardless of format, so they must do their best in good faith to interpret existing laws, to act lawfully, and to act in the public interest.

In 2008 research libraries founded HathiTrust, a digital preservation and access collaboration of over 60 research libraries in the United States, Canada, and Europe. HathiTrust utilizes a shared infrastructure to preserve digital copies of now over 10 million volumes digitized from print. HathiTrust's mission is "to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge."¹ With partners that include libraries in major public and private colleges and universities, independent research libraries such as the Getty and the New York Public Library, and the Library of Congress, and a collection that is increasingly comprehensive, HathiTrust is rapidly becoming a central entity for preservation of and access to library collections. Any given HathiTrust partner library is likely to find more than 50% of its print collection online in HathiTrust. With such a large aggregate collection, the range of works in HathiTrust represents the full spectrum of research library collections, including the copyright status of materials in those collections. Consequently, the HathiTrust libraries must navigate current copyright law to practice responsible stewardship of library collections and to continue their service mission, in the digital realm.

This paper introduces the goals of HathiTrust, describes the scale and scope of the HathiTrust collection and its significance, and discusses how the organization is providing services related to the digital collection, in light of changing conditions for maintenance of the participating libraries' print collections, and, in particular, in the context of the current environment for intellectual property rights.

2. Goals of HathiTrust

Structurally, HathiTrust is not a "trust" in the legal sense of the word, nor is it a corporation or even a non-profit organization. It is a collaborative enterprise of research libraries that depends on funding and in-kind contributions from its members.

The name HathiTrust was chosen to reflect the values of the organization. Hathi (pronounced hah-tee) is the Hindi word for elephant, an animal that symbolically represents memory, wisdom and strength. In concert with its overarching mission, the initial goals set by the HathiTrust partners are:

- To build a reliable and increasingly comprehensive digital archive of library materials converted from print that is co-owned and managed by a number of academic institutions.
- To dramatically improve access to these materials in ways that, first and foremost, meet the needs of the co-owning institutions.
- To help preserve these important human records by creating reliable and accessible electronic representations.
- To stimulate redoubled efforts to coordinate shared storage strategies among libraries, thus reducing long-term capital and operating costs of libraries associated with the storage and care of print collections.
- To create and sustain this "public good" in a way that mitigates the problem of free-riders.

¹ http://www.hathitrust.org/mission_goals

- To create a technical framework that is simultaneously responsive to members through the centralized creation of functionality and sufficiently open to the creation of tools and services not created by the central organization.²

HathiTrust has values of openness and collaboration, and aims to be transparent in its governance and operations. In addition to digital preservation, the organization also aims to provide access to materials to the extent legally permissible.

3. The HathiTrust Collection

The HathiTrust collection has its origins in mass digitization projects conducted in partnership with Google and the Internet Archive, but incorporates much more. HathiTrust brings together collections from many major Google library partners, including the two largest, University of Michigan and University of California, and has the largest collection of items digitized by Google. HathiTrust has also gone a long way towards archiving digital volumes created during Microsoft's Live Search Books 2006-2008 project, volumes which were digitized by the Internet Archive and others. More recently, HathiTrust has focused on incorporating materials that have been locally digitized by the partner libraries. Although HathiTrust content primarily originates from libraries within the United States, the HathiTrust partnership includes international partners such as Biblioteca de la Universidad Complutense de Madrid in Spain, which has contributed a large amount of content, and McGill University in Canada. From its inception in 2008, HathiTrust has grown to include over 10.5 million volumes.

HathiTrust aspires to comprehensiveness and has used mass digitization to accomplish that goal. Consequently, HathiTrust does not have a collection development policy that requires the partners to adhere to any specific subject, language, or content criteria. The HathiTrust partner libraries believe that the value of HathiTrust is in the whole collection, and that this aggregation is reflective of research library collections selected for scholarly value and preserved over time in print by libraries. The aggregate collection also offers the opportunity for differentiation of specific digital collections from the whole, post-digitization, via a layer of services. For example, by aiming for comprehensiveness, HathiTrust is more easily able to offer up sub-collections like English language literature before 1800 or US federal government documents. The vast collection holds the potential to be curated and presented in a multitude of ways using tools available now or developed later. Like the research library collections encompassed within it, HathiTrust serves a broad constituency by incorporating works that did not top the best-seller lists but that serve the "long tail" activities of specialized research and scholarship.

The collection spans the gamut of languages in research libraries: more than 400 languages are currently represented in HathiTrust, of which the highest percentages of volumes are in English, German, French, Spanish, Chinese and Russian. Most languages are present in the collection in smaller percentages, but because the collection is so large, the percentages still represent large numbers of digital volumes, for example Indonesian (33,726), Norwegian (15,429) or Afrikaans (1,053).

² Ibid.

4. Significance of the Collection

HathiTrust serves as shared infrastructure for partner libraries to use in managing their print collections. The significance of a *preserved and dependable* collection of this magnitude is beginning to be appreciated. In early 2011, an OCLC Research study by Malpas reported results of an analysis of the HathiTrust collection relative to the volumes held by US research libraries in print, and found HathiTrust to be increasingly representative of the physical collections in research libraries.³ This holds a number of implications for the libraries in terms of greatly needed understanding of how much of what is held there has been digitized, and how much remains to be digitized; the costs to digitize the remainder cannot be determined unless we know the scale and scope of what is left.

If a given library can possess an understanding of how its particular collection maps to the digitized whole, and can rely on HathiTrust for preservation of digital versions of those books, the library can then make informed decisions about how and where to store its physical book collection, including which print books are essential to keep. When digital books are collaboratively made available, advantages can accrue through collaborative agreements for retention of the physical books, allowing libraries to reduce storage costs in the presence of widely available digital copies.

5. How HathiTrust Provides Services in the Context of the Current Environment for Intellectual Property Rights

The HathiTrust corpus includes millions of works, including both public domain and in-copyright books and serials. HathiTrust can store these works because US law places limitations on the exclusive rights of the rights holders, and those limitations support both fair use and preservation purposes. In order to provide access to the digital volumes in its collection, HathiTrust relies on US and international copyright law and rights determinations for the corresponding print volumes. For example, HathiTrust uses the publication dates and countries of publication in cataloguing records to identify large bodies of public domain works, and in many cases rights holders grant HathiTrust permission to provide open access to materials in the collection. The totality of the HathiTrust strategy can be characterized as a combination of automatic rights determinations, manual rights determinations, permissions and agreements, and legal interpretations.

5.1 Automatic Rights Determinations

HathiTrust's automated rights determination processes identify materials that we can reliably characterize as being in the public domain, either based on US law or common attributes of non-US copyright law. By analysing a number of fixed and free fields in the MARC record, we make a first pass at identifying public domain works, characterizing the remainder as presumptively in copyright. Although we are not able to exhaustively detail the criteria that we consider in making these determinations, several key examples will help illustrate the process. Most US works published in the United States before 1923 are in the public domain worldwide. US law defines the majority of US federal government publications as

³ Constance Malpas, *Cloud-Sourcing Research Collections: Managing Print in the Mass-Digitized Library Environment* (Dublin, Ohio: OCLC Research, 2011), <http://www.oclc.org/research/publication/library/2011/2011-01.pdf>.

public domain. US law also treats non-US works published before 1923 as public domain in the United States; consequently, HathiTrust provides open access to these publications for users coming from network addresses within the United States. To provide access to non-US works for users outside of the United States, HathiTrust generally relies on the Berne Convention⁴ as a framework for decision-making. In many countries, non-US works are only in the public domain 70 years after the death of the author, and because author death date information cannot be reliably inferred in the cataloguing record, we have created a “rolling wall” of 140 years before the current year. (This approach provides us with some protection against assuming a public domain status for a work where the author published a work at a young age and lived for an exceptionally long time). HathiTrust’s automated routines for determining the public domain status of works is published online on the HathiTrust website.⁵ Although the current decision-making framework is focused primarily on US copyright law, Canadian HathiTrust partner libraries have begun discussions to define specific exceptions found in Canadian copyright law.

5.2 Manual Rights Determinations

HathiTrust also uses a carefully defined set of procedures, systems and legal guidance to make manual rights determinations. With generous assistance from the Institute of Museum and Library Services, HathiTrust implemented a Copyright Review Management System (CRMS) in 2008. The design of that system was guided by and continues to be refined by legal scholars. It incorporates strategies such as double-blind review in order to increase the reliability of determinations. The reliability of determinations has been and will continue to be tested against benchmark data (e.g., record analysis by the US Copyright Office). The first CRMS work was focused on books published in the United States between 1923 and 1963; current work is focused on non-US books published in English-language speaking countries and Spain. The Copyright Review Management System is documented online.⁶ Additionally, legal experts may flag individual works for review. All manual decisions override automated decisions, and both sets of decisions are registered in a HathiTrust-maintained Rights Database. The architecture and decision-making related to the Rights Database is also documented online on the HathiTrust website.⁷ More than 100,000 works have been opened using manual determinations; roughly 55% of the works reviewed have been found to have a public domain status.

6. Permissions

Recognizing that many rights holders believe that open, online access to their publications is either part of their mission or in their best interests, HathiTrust supports several strategies for individuals and organizations to open access to their publications. HathiTrust makes a form available online so that the rights holder may convey perpetual and non-exclusive permission for access; this form supports many methods for access, including simple permission without changing the copyright status of work, and application of a Creative Commons license.⁸ Additionally, HathiTrust has negotiated agreements with

⁴ http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html

⁵ http://www.hathitrust.org/bib_rights_determination

⁶ <http://www.lib.umich.edu/grants/crms/>

⁷ http://www.hathitrust.org/rights_database

⁸ http://www.hathitrust.org/permissions_agreement

some rights holders (e.g., Duke University Press) where entire lists of titles are opened with a Creative Commons license, and, in return, HathiTrust provides files and updates to the rights holder. More than 7,000 works have been opened through explicit rights holder permissions.

Using this combination of strategies, HathiTrust has opened more than 3 million works for users at US network addresses, and more than one million works for users worldwide. This represents roughly 30% of the HathiTrust collection. The remainder of the HathiTrust collection, approximately 70% of the works, has either been determined to be in copyright through the CRMS process or is assumed to be in copyright, pending further investigation.

7. Other Lawful Uses of Digital Materials

HathiTrust has used legal guidance to undertake other strategies to provide access to works in the HathiTrust corpus. Rubrics such as fair use in US copyright law (or fair dealing in other regimes) provide a framework for some uses, and HathiTrust supports some of these. In addition, constituencies like the blind or other persons with print disabilities may be served under legal regimes like that in the United States. Similarly, some provisions of US law support HathiTrust's preservation mandate.

Under US copyright law, including the fair use provisions, HathiTrust has developed and provides a powerful discovery mechanism for the entirety of the corpus. Every word and phrase (in hundreds of languages and many character sets) in HathiTrust is indexed and searchable by users worldwide. Where HathiTrust has determined that a work may be made accessible to a given user, the search results provide a significant amount of context, and links are provided to the full text, which can then be read online. In other cases, either in the limited number of instances where we know the work to be in copyright, or where we treat the work as being in copyright in the absence of more reliable information, HathiTrust reports the page numbers and the number of hits per page to the user who conducts a search. This powerful search capability has been extremely helpful to many scholars, as it serves as a master index to a corpus of billions of pages.

Many legal regimes support use of in-copyright works for users with print disabilities. For example, the Chaffee Amendment to US copyright law, Section 121, allows an authorized entity to provide access to works that are protected by copyright to certain users.⁹ In addition, certain uses of in-copyright works to make them available to the blind have been determined to be fair use under US copyright law. The mechanisms HathiTrust has put in place are preliminary, pending the resolution of the legal challenges facing HathiTrust. Currently, using this framework, HathiTrust provides access to millions of works for University of Michigan users certified to have print disabilities. In each case, HathiTrust provides the authenticated user access to the underlying text through a special interface so that the user may use the text with a digital Braille or other reading device. Only digital copies of works that have been determined to be part of Michigan's print collections are included in the service.

Preservation-related provisions in law support other lawful uses of works that are in copyright. In US copyright law, Section 108 supports limited services when an in-copyright work is not available on the market in an unused copy at a reasonable price, and where the library's copy is damaged, deteriorating, lost or stolen.¹⁰ As with services for the print-disabled, the mechanisms HathiTrust has put in place are preliminary, pending the resolution of the legal challenges facing HathiTrust. Currently, at the

⁹ See, for example, <http://www.loc.gov/nls/reference/factsheets/copyright.html>.

¹⁰ See, for example, <http://www.law.cornell.edu/uscode/text/17/108>.

University of Michigan, HathiTrust provides access to certain damaged, deteriorating, lost or stolen works under this interpretation of US law, only to University of Michigan users. These University of Michigan users are not able to download the work in its entirety (i.e., they are currently only able to read the work continuously on the screen or to download one page at a time). No more than one simultaneous user per copy owned may view the work. Each work is clearly marked as being in copyright and the user is notified that access is supported under this interpretation of US copyright law.

As librarians, we must navigate a complex intellectual property rights landscape. Because of the importance and complexity of our work, our University counsels and other legal scholars guide us in making these decisions. Recent work by Peter Jaszi, Jennifer Urban, Pam Samuelson, and other American legal scholars has been helpful, but practical decisions for an organization like HathiTrust are largely untested. We hope, through the processes documented here, to build responsible foundations upon which other uses can be defined.

8. Conclusion

As a digital research library collection unprecedented in size and scope, HathiTrust serves an increasingly pivotal role. HathiTrust has become a vehicle to support end user access to the record of human knowledge and to support the preservation of library collections. Libraries have existed for hundreds of years, each building its distinctive collection with more or less complementarity to other collections. Now, through aggregation, libraries are using HathiTrust to explore questions of bibliographic identification, of collection management, and of copyright determination. Through this collectivity, libraries have begun to make strides in facing the economic and legal challenges inherent in the management and use of digitized library collections.

References

- Berne Convention for the Protection of Literary and Artistic Works,
http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html.
- Cornell University Law School, Legal Information Institute. "17 USC § 108 - Limitations on exclusive rights: Reproduction by libraries and archives," <http://www.law.cornell.edu/uscode/text/17/108>.
- HathiTrust. "Mission and Goals," http://www.hathitrust.org/mission_goals.
- _____. "Bibliographic Rights Determination," http://www.hathitrust.org/bib_rights_determination.
- _____. "Permissions Agreement," http://www.hathitrust.org/permissions_agreement.
- _____. "Rights Database," http://www.hathitrust.org/rights_database.
- Library of Congress, "NLS Factsheets: Copyright Law Amendment, 1996," <http://www.loc.gov/nls/reference/factsheets/copyright.html>.
- Malpas, Constance. 2011. *Cloud-Sourcing Research Collections: Managing Print in the Mass-Digitized, Library Environment*. OCLC Research.
<http://www.oclc.org/research/publication/library/2011/2011-01.pdf>.
- University of Michigan Library. "Copyright Review Management System - IMLS National Leadership Grant," <http://www.lib.umich.edu/grants/crms/>.

The Durationator® Copyright Experiment

Elizabeth Townsend Gard

Tulane University Law School, USA

Abstract

Copyright is a balance between copyright holders and the public, where copyright holders are given a limited monopoly but at its completion, the public is free to use the works as it wishes. But when does a work transition from protected by copyright to its new life in the public domain? That depends on the particular law of the country in which one wants to use the work. This becomes fairly overwhelming in a digital age. For the last five years, we have been creating the Durationator, a software tool that allows users around the world to input specific information about a particular cultural work and obtain legal information regarding the copyright status of the work—for the U.S., for specific regions, for the whole world. As more and more old works are saved, preserved and made available in a digital context, the need for such a tool becomes more urgent. Our project has researched and is in the process of coding the copyright law (in terms of duration) for every country in the world, including dependencies.

Author

Dr. Elizabeth Townsend Gard is an associate professor of law at Tulane University Law School, co-director and co-founder of Tulane's Center for Intellectual Property Law & Culture, and director and co-inventor of the Durationator® Copyright Experiment, a software program that determines the worldwide copyright status of every kind of cultural work. Since 2004, she has been a non-resident fellow at the Stanford Law School Center for Internet and Society. She, along with her husband, have begun the Copyright Research Lab and Help Center, which will offer low-cost self-help solutions to copyright questions

1. Introduction

One of the greatest challenges facing preservation and access research and development is copyright law: when are works in the public domain, when do library exceptions or fair use apply, when does state law apply (sound recordings, for example), how are foreign works to be treated, and which works can one post on the Internet without facing liability. These are just a few of the questions plaguing librarians, artists, scholars, teachers, corporations, the content industry, digitizers, students, hobbyists—everyone in a digital age.

The Durationator® Copyright Experiment tackles the question of when, how, and in what circumstances people (librarians, scholars, filmmakers, teachers, hobbyists, digitizers) can use cultural works.¹ We see the need for access to accurate, accessible, quick, and low/no cost solutions as key to work in the arts and humanities, to preserving the culture, and to bringing the old into the Internet Age, e.g., digitization. In the last five years, at Tulane Law School, we have devised procedures for determining the copyright status of a work worldwide. We have researched and coded the copyright laws of every country in the world. To date, only copyright experts in the field have seen the work. Starting in the Summer 2012, we began branching out to individuals and our strategic research partners to understand

¹ Tulane Law Students made a two-minute video about our project available at our website www.durationator.com.

how our raw research can be applied within specific settings—libraries, individual scholars, artists and filmmakers, publishers, digitizers, lawyers, and the corporate content industry.

The research for the project has been our main focus, and has been arduous, both in depth and scope. Many unexpected significant research questions had to be answered and that research then often had to be replicated in 200+ countries in the world. The research is never-ending but we find that we have developed a system to tackle these hard questions from which others shy away. Our project does not directly digitize works, nor does it create a database of public domain images. Rather, the Durationator® Copyright Experiment provides legal information for specific queries from users, who bring information about a particular work to the software tool.

The project began as a quest to understand U.S. law for domestic and foreign works, but it quickly broadened to include current law for every country, including dependencies in the world. Copyright may be territorial, according to the law, but the reality of our world is that we are all engaged in a global world. In the end, we hope our tool could be used in any instance to determine the copyright status of a cultural work, whether for a local or global use.

2. Copyright and Code

In January 2012, Justice Breyer, in his dissent of *Golan v. Holder* referenced my research and work with the Durationator® Copyright Experiment at Tulane University Law School. In explaining why copyright restoration of foreign works that had fallen into the U.S. was problematic, Breyer wrote, "...the statute's technical requirements make it very difficult to establish whether a work has had its copyright restored by the statute. Gard, In the Trenches with §104A: An Evaluation of the Parties' Arguments in *Golan v. Holder* as It Heads to the Supreme Court, 64 Vand. L. Rev. En Banc 199, 216–220 (2011) (describing difficulties encountered in compiling the information necessary to create an online tool to determine whether the statute applies in any given case)."² Personally, being cited by the U.S. Supreme Court is an important moment in one's career as a legal scholar. For me, it was particularly sweet, because our hard work here at Tulane University had been recognized in the most significant forum in law—the Durationator® Copyright Experiment had arrived.

I have been researching questions of copyright within a practical, real world scenario for seven years.³ I have come to believe that only code can make law accessible, and therefore, allow the average

² *Golan v. Holder*, 132 S.Ct. 873 (2012).

³ Published papers include Elizabeth Townsend Gard, "A Tale of Two Ginsburgs: Eldred v. Ashcroft and *Golan v. Holder* (DePaul L.R., forthcoming); Elizabeth Townsend Gard and Erin Anapol, *Federalizing Pre-1972 Sound Recordings: Two Proposals*, co-authored with Erin Anapol (Tulane J. of Technology and IP, forthcoming); Elizabeth Townsend Gard, *In the Trenches with Section 104A: An Evaluation of the Parties' Arguments in Golan v. Holder as It Heads to the U.S. Supreme Court*, Vanderbilt Law Review (invited, 64 VAND. L. REV. EN BANC 199 (2011); Elizabeth Townsend Gard, "The Making of the Durationator®: An Unexpected Journey into Entrepreneurship," book chapter in *Entrepreneurship and Innovation in Evolving Economies: The Role Of Law*, ed. Megan Carpenter (Edward Elgar Publishing, 2012, forthcoming); Elizabeth Townsend Gard, and Copyright Class 2011, Reply Comment, Pre-1972 Sound Recordings, U.S. Copyright Office (created collaboratively with the 2011 Copyright Class), April 13, 2011, available at <http://www.copyright.gov/docs/sound/comments/reply/041311elizabeth-townsend-gard.pdf> (*cited and proposal partially adopted by Copyright Office report), Elizabeth Townsend Gard, "Copyright Law v. Trade Policy: Understanding the Golan Battle within the Tenth Circuit," *Columbia Journal for Law and the Arts* 34, no. 2 (Winter 2011): 131-199; W. Ron Gard and Elizabeth Townsend Gard, "Marked by Modernism: Reconfiguring the 'Traditional Contours of Copyright Law' for the Twenty-First Century," in *Modernism and Copyright*, ed. Paul Saint-

user—everyone—to understand the choices we make when we post a family photograph on Facebook, choose to include a map in our scholarly publication, or decide which music to include in a documentary film. Our experiment at Tulane University attempts to use code to make available highly technical and difficult copyright law.

Can code make law accessible? In the 21st century, many laws have become so complex, particularly in an Internet age of multiple jurisdictions, that few average people can understand our laws. Lawyers themselves have trouble sorting through which laws apply, and if you have questions regarding copyright and posting works on the Internet, one could conceivably need to consult 220 different laws, for example, just to assure a work is in the public domain. Humans cannot perform this task. Code must come to our aid.

The Durationator® Copyright Experiment began as a research problem to a simple question: **can one determine the copyright status of a cultural work in the context of posting the work on the Internet?** Since no “Internet” copyright law exists, this meant to determine whether a work—a poem, novel, film, photograph, computer software—was either protected by copyright or free for all to use in the public domain—one must determine country-by-country the copyright status of that particular work. While the Berne Convention works to harmonize copyright laws, in fact, many differences appear, and some countries, including Russia, the U.K. and especially the U.S. are particularly complicated. After years of research and coding, we now know the copyright status of any specific work in each individual country of the world.⁴

3. Significance

3.1. The Durationator® For All Users of Culture

We believe that users of all kinds will benefit from legal information delivered in a form that is accurate, accessible, immediate, and at low/no cost. In short, the software was designed for my graduate self—an American worried about which British works I could use from a Canadian archive. It turned out that my worries were the worries of the world—from the Google Book Project and HathiTrust, both engaged in lawsuits over copyright infringement to the independent author writing a book on perfume history to the Disney studio engaged in negotiations over the rights of *Bambi*.⁵

We have worked with a number of different individuals and groups over the years—and we have found that whether it is Electronic Frontier Foundation, a movie studio, University of Michigan, an

Amour (Oxford University Press, 2010), 155-172 (invited piece and published); W. Ron Gard and Elizabeth Townsend Gard, “The Present (User-Generated Crisis) is the Past (1909 Copyright Act): An Essay Theorizing the “Traditional Contours of Copyright” Language,” *Cardozo Arts And Entertainment Law Journal* 28 (2011): 455-499; Elizabeth Townsend Gard, *Introduction to Shirley Millard’s I Saw Them Die* (1936, reprinted Quid Pro Books, 2011) (invited; introduction to public domain work in the U.S.); and Elizabeth Townsend Gard, “Unpublished Work and the Public Domain: The Opening of a New Frontier,” *Journal of the Copyright Society of the U.S.A.* 54 (Winter 2007).

⁴ We also know the copyright status of any work within historical time for about ten jurisdictions. That is, if you were living in 1799 in the U.S., would that particular painting, for example, have been under copyright? In 1840? 1910? We found the historical paths, as we call it, as important as current law, and have spent a good deal of time with Russia, Germany, Israel, Japan, China, France, the United Kingdom, the Berne Convention itself, and are working on other countries. This requires an understanding of case law, statutes, amendments, historical works, and custom.

⁵ *The Authors Guild v. Google*, 770 F. Supp. 2d 666. *Authors Guild v. HathiTrust*, Complaint SDNY. *Twin Books Corp. v. Walt Disney Co.*, 83 F. 3d 1162.

independent scholar, a book publisher, or a graduate student, the Amistad Research Collection—the same questions arise. All of these groups want to know the legal status of the work—what they can and can't do with a particular work. Their expertise in copyright may vary. But their needs and questions are consistent. To that end, we have developed a couple of strategies. First, the Durationator® itself will eventually be available to the public, with a no/low-cost version available. In the meantime, we are developing a business model that combines DIY Copyright Coaching (legal information) with one-on-one legal advice with an attorney. Finally, we will provide Durationator® Reports, Auditing Services, and Consulting as full-service legal advice as well. The research will continue to be done at Tulane Law School. The spinout services and products will be housed at Limited Times, L.L.C.

3.2 The Durationator in Libraries and Digitization Projects

3.2.1 Libraries

We believe the Durationator® Copyright Experiment will provide a valuable solution to daily copyright problems in the library by helping to determine if a particular item can be added to a digital archive, used by a researcher with or without restrictions, or transmitted electronically to remote users. In the age of electronic communication, users are requesting to have materials available electronically at an ever-increasing rate. Most librarians however, are reluctant to reproduce material that could still be under copyright protection. The Durationator® Copyright Experiment will provide librarians with a simple method for determining the copyright status of the research materials their users need most and make digitizing unique or historic collections far simpler.

Additionally, a strange disconnect occurs. Libraries digitize and post works, and look to the copyright questions for their own liability issues. We have seen broad categories being applied, a generalization. But for the user coming to that work—to include in a book, a movie, on a website—we must actually *know* the copyright status. We are not, as users, protected by the library's assertion "no known copyright restrictions." The Durationator can bridge the gap. It can take the information provided by the library and provide a tool for users to do a specific legal search related to that specific item. More certainty.

We also see the need for the Durationator® with archival materials, especially in their non-digitized state. Making use of valuable primary source materials in library special collections will be far simpler with the help of the Durationator®. A Durationator® search results report could be attached to the record of items in a library's special collections, allowing users to immediately know which works they could freely use in their research and which would need permission. Upon seeing the attached search results, researchers would also learn about the Durationator® and would know to use the service if they had further questions about additional materials they encounter throughout the research process.

In Interlibrary Loan, libraries could perform a Durationator® search on any item lent to another library before shipping the item through the mail.⁶ If the work is no longer under copyright protection and is often requested, digitizing the work and sending a digital copy in the future will save time, money, and the environment by cutting down on shipping costs and eliminating the risk of damage in transit. For academic libraries, copyright issues are a major concern for administering electronic course reserves

⁶ See the recent decision in *Cambridge University v. Becker* (delivered by the district court on May 11, 2012) regarding interlibrary loans at Georgia State. The 350 page opinion brings home the issue of copyright's role in libraries' daily activities.

systems. With the help of the Durationator®, librarians and professors could identify which works require permission and take affirmative steps to ensure that copyright law is observed while still making the work available for students.

Finally, the Durationator® Copyright Experiment will contribute to raising awareness of copyright issues for libraries and their users and allow historians, hobbyists, educators, and researchers to feel confident that they are complying with the law in their own work. By providing access to Durationator® searches for library materials, libraries can promote the use of their collections confidently and responsibly, making valuable research materials more accessible to all.

The Durationator® Copyright Experiment has formed strategic research partnerships with Louisiana libraries to assist in testing and adapting the product to serve the many needs of the library community—through LOUIS, the Louisiana Library Network, MediaNOLA, and the Amistad Research Collection at Tulane University. In addition to the librarians with copyright expertise who are serving on the Copyright Advisory Board (Kenneth Crews and Peter Hirtle), Tulane law student and former Access Services librarian Kathryn Munson has also been recruited to the product development team as the Director of Library Research to meet the needs of this unique service group, and will serve as our Director of Library Research and Outreach.

3.2.2 Digitizing Projects

Digitization projects—in libraries, by corporations like Google, and by hobbyists—abound. Moreover, litigation regarding copyright infringement over digitization projects raises the profile of the need to determine properly the copyright status of the many millions of works being digitized.

We think the Durationator® Copyright Experiment could assist with digitization projects, both large and small scale. We also realize that just because a digitization project has determined a work is in the public domain—say, as part of the proposed Google Book settlement—does not mean that a user can depend or receive benefit from the safe harbour provisions that Google receives. This means that a user using a Google book *should run their own search*. If accepted, Google will not be penalized if they get the answer wrong, but users using a particular work, could be.

We also believe there are more works to be digitized beyond books. To date, the majority of the projects have focused on either books (the easiest category to determine) or pre-1923 works (again an easy category). We want to encourage the many other layers of culture that remain underutilized because of the difficulty of determining their copyright status. We’ve worked with Quid Pro Books to help uncover previously unknown public domain books, including the republication of Shirley Millard’s *I Saw The Die*. We are also now working with University of Michigan as an independent third-party auditor of their large-scale project to determine the copyright status of thousands and thousands of books. University of Michigan’s library has become one of our Strategic Research Partners in determining the copyright status of foreign works, and also serving as an outside auditing system for their work.

3.3 Understanding Current Litigation and Copyright

Three major cases have involved cultural works held in libraries and their availability and access in a digital context: Google Books and the Proposed Settlement (that was rejected by Judge Chin), the Author’s Guild v. HathiTrust litigation over making available “orphan works”, works still under copyright but whose authors cannot be located to give permission, and Cambridge University v. Becker, regarding course materials in a digital context at Georgia State. Each of these cases has at their core what

libraries can and cannot legally do with works, and this depends on copyright status of each individual work. The Durationator® Copyright Experiment has watched this litigation storm with great interest. If the Google Book settlement is approved, Google will have a safe harbour on “mistakes” they make in determining the copyright status of works (labelling works in the public domain when they are actually still under copyright). However, users of the Google books will not have the same protection, nor will Google’s determination of the copyright status of the work be good for anyone but Google. The settlement does not allow users to legally depend on the analysis performed by Google. We think the Durationator® will be needed more than ever if the settlement is approved. What is clear from this litigation is that libraries, scholars, and others in the humanities are not immune from litigation, and a tool like the Durationator® is needed more than ever.

4. A Personal Context for Creating the Durationator® Copyright Experiment

My own journey with copyright began in 1987, when I first encountered what would become my dissertation subject. While taking an undergraduate course in the Culture of War, I was assigned the Great War diaries of British writer, Vera Brittain (1893-1970). The next week, another student presented Brittain’s memoir on the same time period, and I remember the stark differences in interpreting the same events, with the student actually accusing me of getting the events all wrong. I was encouraged by the professor (later my dissertation chair) to write a seminar paper on the two works, my undergraduate honours thesis on Brittain’s larger transformation, my master’s thesis on Brittain’s interwar years, and my dissertation as a comparative generational biography, with Brittain as the focus but in comparison to other men and women writers of her generation.

The broadening of my dissertation from merely a study in Brittain was due to copyright questions arising from the Brittain papers, and whether Brittain’s literary executor (also writing a biography at the time) would give permission to publish. Diversifying solved the problem, and also proved a more interesting project. But I found myself exposed to the uncertainty of copyright and how it affected my daily work. Could I use a 1917 unpublished letter?⁷ How was this different than quoting from the published version?⁸ Who did I need to ask permission from to use a particular photograph, clearly not taken by the Brittain family, and did one always have to ask permission or could I actually rely on fair use? What did it mean when a work went into the public domain, and how would one determine if that had occurred? I loved all of the materials I was encountering, but I felt unsure about what works I would actually be able to use in a publication, and what restrictions I might face from a literary executor.

At the Centenary conference for Vera Brittain, where I presented my work on gender and generational theory, I began hearing more and more stories from Brittain scholars of the difficulty of getting permission to publish from the current literary executor, who was writing his own biography of Brittain. The warnings would have scared any young graduate student. I also started to hear stories from other scholars in history that had abandoned projects for fear of getting entangled in uncertain legal status

⁷ The unpublished portions of the letter are in the public domain in the U.S. However, the published portions of the same letter are under copyright through December 31, 2047. 17 U.S.C. 303(a). At the time I was working on the project, however, the entire letter was under copyright in the U.S.

⁸ It turned out that the published memoir was in the public domain in the U.S. while I was a graduate student, but was “restored” by a complicated part of the copyright law in 1996. *Golan v. Holder*, the recent U.S. Supreme court case addressed this issue. I have spent a good deal of my career focused on “restored” foreign works, both within my research and the implementation of the statute in the Durationator®.

of works, and having nowhere to turn. Then, only in my early-twenties, I turned to my father, who, like all of the men in my family, was a lawyer. He took me to a law library in Los Angeles, which opened a whole new world, changing my path forever. For it was here I found and began to research the legal doctrine of fair use, copyright term, and other elements that have now come to dominate my life.

Upon completing my doctorate in May 1998, I enrolled in law school the following fall.⁹ I wanted to find answers to the questions that haunted not only my work but also others' I had encountered during my graduate work. It was also at this time, in the mid-1990s, that faculty began asking themselves what could they post and include on this new idea of a "web page" for courses, and like the materials of scholarship itself, no one seemed to have reliable legal answers on which to depend. I set out to understand copyright within an academic setting, and forge my own path as an advocate for scholars and teachers.

After completing my law degrees at University of Arizona, I was offered a Leverhulme Fellowship at the London School of Economics, to pursue research on copyright issues affecting academia, and to teach copyright (UK and International). There, I began to realize that the issues I had seen in my doctoral work had a distinctly important international component, and was broader than merely my own struggles as an academic writing 20th century biographies. The whole world seemed to have the same questions: when could they legally use a particular work? By 2005, the world, once analogue, had become digital, and copyright law, which was structured on a 19th concept of territory, now had to adapt those concepts to an instantaneously global world. My small questions of which works could she use in her dissertation now had global dimensions.

The postdoc led to a Visiting Assistant Professor position at Seattle University in 2006, and a tenure-track position at Tulane University Law School in the Fall 2007. It was in the Summer 2007 that the Durationator® Copyright Experiment was born. Two events occurred. A rising 2L law student with computer coding experience came to work for me, and an unsolicited email arrived asking if it was legal to use a Vera Brittain poem from 1918 as lyrics to a new musical composition—that is, was that poem in the public domain? As part of my research and job talk that year, I had been working on a piece on copyright and the public domain, and in particular the status of unpublished works. I knew the answer would be complicated. I didn't realize how complicated. After a number of months of research, I had the answer. I also realized that the system in place made it nearly impossible for anyone—trained in law or otherwise—to actually determine the copyright status of works. The idea of a software tool had been born. I always knew that I, along with my spouse (also a J.D./Ph.D.), had wanted to start a clinic or outreach program to assist scholars, students and teachers. Now, I wanted to make the copyright status of every work ever created—anywhere in the world—available.

Every work has a particular term of protection, and then after that term expires, all can use the work without permission. The work is in the public domain. The legal question I sought to answer: when did that occur—in the U.S., in the U.K., by posting it on the Internet, by disseminating it in Spanish-speaking countries, by making it available to the world? I knew it would not be an easy question—my brush with comparative copyright in the UK had taught me quickly how complicated international copyright issues could get. But with my 2L research assistant, we set out to research and code the world's copyright laws, and to answer the question: when does any particular work come into the public domain? We naively

⁹ I actually hid this information from my doctoral chair until I completed my first year. It was the biggest secret I ever kept. He had been so supportive since my late teens and my first discovery of Brittain. I didn't want to disappoint him in my divergent path, but I also felt compelled to know the answers in my quest.

thought it would take a summer. After five years, we have finished Phase One of the project, and we are now on to Phase Two, making the information and access to help available to the public.

Along the way, one more important element occurred...I found an intellectual community that supported and cared about the work that I was doing. It has been a whirlwind of exciting activities over the last five years—far more than I ever dreamed in my pre-tenure life. I've met, shown my work, shared meals, and been encouraged to continue the Durationator® Copyright Experiment by individuals I never dreamed I would even meet—David Nimmer (Nimmer on Copyright), Bill Patry (Patry on Copyright, Google Senior Copyright Counsel), Paul Goldstein (Stanford), Jule Sigall (Microsoft Senior Copyright Counsel), Peter Jaszi (American University), Pam Samuelson (U.C. Berkeley), David Carson (U.S. Copyright Office), Diane Zimmerman (NYU), Graeme Dinwoodie (Oxford), Daniel Gervais (Vanderbilt), Tony Reese (Irvine), Kenneth Crews (Columbia), to name just a few. Together, this group comprises some of the smartest folks on copyright in the U.S., and in particular duration, the public domain, and international issues. Each of them have sat with me, talked with me, and even reviewed what we are doing on the project. They all have not only been encouraging, but they most have agreed to act as our Copyright Advisory Board, which includes being part of our alpha prototype testers. This is what makes our project so unique. It is an academic working to find answers for other academics and productive users of cultural works researched by a law professor in a culture where other law professors intellectually support the work at hand. It has been the most amazing experience.

For all of this, I have stayed true to the project, and thanks to Tulane University's support, have been free to develop the mission and goals free from outside restraint or pressure. I want the software to help my old self—the graduate student wanting to know which works I had to ask permission from the literary executor, which works might qualify for fair use, and which were in the public domain. I don't want others to have to go to law school, spend five years and thousands of student research hours, and over \$100,000 in university funding to find the answer. I want DIY services that will help train those interested in self-help, while at the same time also providing them resources to legal advice. I want, in the end, to help the world determine the legal status of our cultural treasures.

The project has benefited greatly by the dedication and devotion of so many law students over the years. They have been amazing, individually and as a group. The work, ideas, and intellectual decisions they have all made make the software that much better. They have been my colleagues, and I have watched each of them grow in tremendous ways. It has been one of the great joys of the project—to see their leadership and research skills grow, and then to see them transition into their own careers and interests. We have had over three dozen students work on the project—and I've seen them grow and develop as well. It has been one of the great blessings of the project.

One more component significantly added to the support and success of the project: my spouse, W.R. Gard. First, he was willing to move many times, even when it was not in his best interest and his career—a trailing spouse is never easy. For that I am insanely grateful. The opportunity to work in a comparative and internationally focused law school at Tulane Law School presented itself to me, and he was willing to take a lesser position at a neighbouring school, instead of going on the market himself. But many spouses sacrifice. It was during dinner conversations that the essence of the project developed, and our method for figuring out how to move through the system took shape—for it was his theoretical work that changed the way we move through the U.S. law. We have co-authored a number of papers on this topic, but essentially, it was only in applying a Marxist reading comparing the turn into the twentieth century with the turn into the twenty-first century in both the culture and the law protecting culture that we came to realize key elements of the system. His theoretical work, his own development of a theory of

reading, has deeply influenced the shape and success of our project. Finally, he has been willing to take on significant responsibilities, including becoming the sole managing-member of our Tulane spin-out company, Limited Times, LLC.

Part of Ron's work focuses on the development of the corporation, and how this influences who we are as people—how the 19th century corporate form, the 20th century corporate form, and the 21st century corporate form deeply affects how we live, love, and physically live in the world. His doctoral and post-doctoral work focuses on developing a theory of reading the corporation through literature and film, using human geography, Marxism, and literary theory. Part of my own struggle with the idea of a spinout company was where we fit within the larger world, and how we stay true to our project. Ron has taken on that struggle, both in form and substance. The LLC, as we envision it now, will remain a closely held corporation, so that we can control its meaning and message. We will offer what we yearned for ourselves so long ago as graduate students—a place where we could learn how to understand the laws ourselves, and empower us to make thoughtful choices. This is what Limited Times LLC will do, and again, thanks to Ron, we will be able to see our vision come true. We want a project that can sustain itself financially, making sure that it can continue to update the technology and the law of every country in the world.

5. Major Research Questions Underlying the Durationator® Copyright Experiment

In developing the software, we have had many, many research hurdles that often have turned out to be very, very, very difficult and time-intensive queries—whole projects onto themselves. Many of these have become paper(s) in their own right, and we will continue to expand the research necessary to understand the state of duration and access in copyright law worldwide. In short, copyright duration is complicated on many levels.

First, within the U.S. context, the records needed to determine the copyright status of works published before 1978, until recently had not been digitized, and therefore were inaccessible to most. Even if one could get their hands on a copy of the Copyright of Catalogue Entries in one of the regional depositories, the extensive records are not intuitive, and their organization changes over the years. We worked a great deal with the original records to try to understand how to determine the copyright status of works, and where the average user would have trouble figuring this out. Now that they are digitized, one has access, but because all but the book records remain in scanned form only and not in a searchable database, their level of inaccessibility remains. I was told by someone at Google that they were working on the problem, but with no known date of releasing the records in database (XML) form, as they did with the book records. This is a very serious problem—as these records tell you whether a work was registered or renewed, was published or unpublished, foreign or domestic, and under what category they were registered—all valuable information needed to determine the copyright status of pre-1978 works in the 21st century.

We also encountered problems with legal questions regarding publication, government employment records, sound recordings, the status of state government works, and foreign laws necessary for determining the copyright status within the U.S. Each topic has turned into a major research project (involving many students) and eventually into research paper(s) the results of which are coded and incorporated into the Durationator® Copyright Experiment.

Our greatest hurdle was Section 104A, and it is this work that Justice Breyer referred to in his dissent in *Golan*. Section 104A restored copyright to foreign works that had fallen into the public domain before their copyright had expired in their home country. How the amendment went about doing this was

problematic on many levels, and to that end, I've been writing in one way or another on this topic for the last three years.¹⁰ Moreover, we had to research each country of the world to determine the copyright status of their laws at the time of restoration (which varies depending on the country), which itself proved very problematic, as most countries have altered their laws in the last ten years, and moreover, the old laws are not very accessible as they are generally from a pre-Internet era. I have had three students over three summers devoted to this project, and we believe, that by January 2013, we will have finally the data to complete our Section 104A path.

But Section 104A will surely be matched by our newest project—the rule of the shorter term. This is a complicated part of the Berne Convention that states that a term of copyright can be shorter if the country of origin's term is shorter than the country's term where protection is being sought. The problems are many. First, a country must elect not to adopt Rule of the Shorter (RST) term. Second, Berne may or may not be self-executing, and therefore, some countries may explicitly not adopt RST, but it would only apply if Berne were self-executing. Third, Berne has a component that exempts RST if there is a bilateral or other treaty in place that speaks to the issue. And so, as we have done before, we must now determine all of these elements for each country in the world. I began to see the problem when I was training a law student new to the project. We were doing research on Zimbabwe, and we got to the question regarding Rule of the Shorter term. No mention was made in Zimbabwe's copyright law. The next question we faced was whether Berne was self-executing or required implementing legislation. We soon realized that whether Article 7 of the Berne Convention applied in any particular situation was very complicated and required extensive research. We think this will be our larger Summer 2013 project.

Foreign laws are not the only challenges we faced. Pre-1972 sound recordings are covered by states rather than federal law. We have worked extensively on mapping each state's laws, and also contributing our comments and suggestions to the U.S. Copyright Office's call for comments and replies in the proposition of federalizing pre-1972 sound recordings. State government works are also complicated, and we are beginning to research how to determine when a state or local government is asserting copyright. Finally, federal government works are no picnic either, requiring a great deal of research. While federal government works are in the public domain, determining what counts as a federal government work involves questions of whether a work falls under the scope of a government employee's employment, among other issues. It's complicated.

The project has also worked on unpublished works—the research in fact is where the project began, and we have now begun working on implementing our research in a practical setting, using the MediaNOLA and courses at Tulane University implementing archival materials as our laboratory. What information does someone using archival materials have at hand, and how can we determine the copyright status using those resources? We are trying to make our research as accessible and error-proof as possible.

We have many other projects—one on social media and copyright, another on the U.S. Holocaust museum and its agreements with museums around the world, historical case studies of copyright in Germany, Israel, China, United Kingdom, Australia, and Japan, and a number of studies focusing on the

¹⁰ The majority of Dr. Townsend Gard's research work has focused on Section 104A, the restoration of foreign works. The first piece looked at the mechanics of the amendment; the second piece (invited by Vanderbilt) looked at the briefs leading up to the oral arguments for *Golan v. Holder*; the third piece analysed the lack of First Amendment analysis in what was billed as First Amendment cases (*Golan v. Holder*, and also a patent case); the fourth piece compares Ginsburg's outcome in *Eldred* with her writing in *Golan* and asks where do we go from here; and the final (hopefully) piece is an invited piece from Franklin Pierce on the relationship of the treaty clause and the IP clause after *Golan v. Holder*.

1909 Copyright Act. All of these projects, and future projects, hope to add to the knowledge base contained within the Durationator® Copyright Experiment.

6. Our Partners and Advisory Board

6.1 Copyright Advisory Board

The board functions to assist in advice with complex research questions (i.e., Rule of the Shorter term interpretation), testing of the alpha prototype, and in particular, reviewing our plans for future research projects. They will also assist in helping us strategize on the final plan for dissemination and sustainability. All of the following IP scholars and practitioners are familiar with our work, and in particular, the Durationator® Copyright Experiment. Some have been involved more than others. Each have seen the Durationator® demonstrated.

- Pamela Samuelson, Professor of Information; Faculty Director, Berkeley Center for Law & Technology; Richard M. Sherman Distinguished Professor of Law
- Peter Jaszi, Professor of Law, Faculty Director of the Glushko-Samuelson Intellectual Property Clinic
- David Nimmer (Nimmer on Copyright)
- Graeme Dinwoodie, Professor of Intellectual Property and Information Technology Law
- Daniel Gervais, FedEx Research Professor of Law
- Peter Hirtle, Senior Policy Advisor, Cornell University Library
- Kenneth Crews, Director, Columbia Copyright Advisory Office
- Tony Falzone, formerly Stanford University Fair Use Project, now counsel for Pinterest
- Julie Samuels, Electronic Frontier Foundation
- Glynn Lunney, Tulane University
- Alan Childress, Quid Pro Books
- Jule Sigall, Microsoft
- Deborah Gerhardt, University of North Carolina School of Law

6.2 Strategic Advising Partners

Over the years, we have worked directly (EFF, Stanford Fair Use) or had discussions about the Durationator® with the following institutions, and we would include them in our discussions on how to make the software accessible and useable to the general public.

- University of Michigan, Copyright Review Management System – World Project
- Stanford Law School Fair Use Project
- LOUIS: Louisiana Library Network
- Amistad Research Center, Tulane University
- Jazz Archive, Tulane University
- Newcomb Special Collections, Tulane University
- Louisiana Special Collections, Tulane University
- Konomark and Dr. Eric Johnson, University of North Dakota
- MediaNOLA and Dr. Vicki Mayer
- Quid Pro books and Dr. Alan Childress

Our work at Tulane and Media NOLA focuses on how to communicate what information is necessary to users to determine the copyright status of works—how does one quickly educate a class in architecture on how to use and determine the copyright status of works in a particular architecture archive? Our work with Quid Pro Books assists Dr. Childress in providing legal information on specific works he is interested in republishing. We are testing out our auditing services with the University of Michigan, and testing our software with LOUIS. In the next year, we are actively looking for additional strategic research partners as we transform our experiment into usable tools and forms. In particular, we would like to find partners outside of the United States to begin understanding the needs outside of the U.S. We have learned so much from questions from individuals and institutions, and we look forward to expanding our scope and learning a great deal more.

6.3 Testing the Delivery of Information

We are now entering into a new phase—how are we going to deliver the information provided by the research and code we have created? To this end, we are working on a number of alternatives, the successful ones that will be available at Limited Times in the near future. We have come to see our role as providing legal information and support to individuals and institutions struggling with copyright questions, and when necessary providing resources for legal advice. Even this required a great deal of legal research—to understand the line between legal information and legal advice (creating an attorney-client relationship), particularly after Legal Zoom and the new development of cloud lawyering sites. We are working on a model that meets the requirements of laws currently, and takes advantage of the new ideas about delivering legal information and legal advice in the digital era.

7. What We Have Learned – Our Journey Ahead

The journey so far has been exhilarating. I never imagined my copyright problems as a graduate student were also the problems of so many around the world. I also never imagined that I would lead a team of three dozen students into complex research and build a practical tool. I never imaged that I would be involved with so many people doing so many interesting projects. I never thought I would be starting a company with my spouse, focused on our passion of making cultural works more accessible. I certainly never even dreamed I would be cited in a Supreme Court opinion (even if it was the dissent). I love the research and work in all its complexities. And now that we are connecting to others and trying to figure out how to help them with their version of the same problem, it has become all that more interesting.

The Intangible and the Digital

Participatory Media Production and Local Cultural Property Rights Discourse

Kate Hennessy

School of Interactive Arts and Technology, Simon Fraser University, Canada, <http://hennessy.iat.sfu.ca/>

Abstract

The 2003 UNESCO Intangible Heritage Convention specifies that communities are to be full partners in safeguarding efforts. Yet the notion of safeguarding has been complicated by the politics and mechanisms of digital circulation. Based on fieldwork in British Columbia and Thailand, I show that participatory productions of multimedia aimed at documenting, transmitting, and revitalizing intangible heritage are productive spaces in which local cultural property rights discourses are initiated and articulated. I argue that digital heritage initiatives can support decision making about the circulation—or restriction—of heritage while drawing attention to the complexities of safeguarding in the digital age.

Author

Kate Hennessy is an Assistant Professor specializing in Media at Simon Fraser University's School of Interactive Arts and Technology (SIAT). As the director of the Making Culture Lab at SIAT, her research explores the role of digital technology in the documentation and safeguarding of cultural heritage, and its representation and exhibition in new forms. She has been a lecturer in the Sirindhorn Anthropology Centre's annual Museums and Intangible Heritage Field School in Lamphun, Thailand, since 2009.

1. Introduction

In the last decade, I have worked on applied projects in visual and media anthropology in Canada and Thailand. In the course of this work, I have come to view participatory media production as central to ethical digital documentation and representation of culture, languages, and heritage. In this same period of time, UNESCO produced the *Convention for the Safeguarding of the Intangible Heritage*¹ and the *Charter on the Preservation of the Digital Heritage*.² Both of these policy documents have become important for thinking through a range of local practices and approaches to documenting and representing intangible heritage in the name of transmission to future generations. These documents, and my fieldwork experiences, have for me also highlighted local approaches to ownership and control of cultural heritage and its digital representation—what I discuss in this paper as *local cultural property rights discourse*—as central to the project of safeguarding.³

For example, between 2004 and 2007, I worked in collaboration with members of the Doig River First Nation, a Dane-zaa community in northeastern British Columbia, and a team comprised of folklorists, anthropologists, linguists, and interactive media specialists, to produce a Virtual Museum of

¹ UNESCO, *Convention for the Safeguarding of the Intangible Heritage*, 2003

² UNESCO, *Charter on the Preservation of the Digital Heritage*, 2003

³ This paper summarizes arguments from a forthcoming article: Kate Hennessy, "Cultural Heritage on the Web: Applied Digital Visual Anthropology and Local Cultural Property Rights Discourse," *International Journal of Cultural Property* (forthcoming).

Canada exhibit called *Dane Wajich—Dane-zaa Stories and Songs: Dreamers and the Land*.⁴ This virtual exhibit features oral narratives and song traditions relating to the history of Dane-zaa dreamers, also known as prophets,⁵ and a contemporary history of the Doig River community and territory as they negotiate their Aboriginal and treaty rights. In the course of producing this exhibit, which involved central participation of youth, elders, and community leaders, significant conversations emerged around the ownership and circulation of documentation of intangible heritage. The ensuing negotiations over intellectual property rights relating to archival documentation of intangible heritage, including what could be shared over the internet, and what should be restricted as private knowledge, represented an important process of articulating local cultural property rights that shaped the content of the virtual exhibit. Safeguarding, in this case, included keeping some elements of intangible heritage documentation out of the public domain.

I have also seen these dynamics echoed in fieldwork in northern Thailand. Between 2009 and 2011, I worked as a lecturer and resource person in the Sirindhorn Anthropology Centre's Intangible Heritage and Museums Field School in the northern province of Lamphun.⁶ In two of these field school seasons, I collaborated with students and members of the Buddhist temple community of Wat Pratupa to identify and document elements of endangered intangible heritage. We observed that at Wat Pratupa, digital media and internet-based circulation of documentation have been central in local efforts to safeguard local cultural practices. Like community-based media projects at the Doig River First Nation, these practices have also opened up spaces for the negotiations of local approaches to sharing, and to the articulation of local cultural property rights. However, where members of the Doig River community opted to keep sensitive heritage off the web, members of the Wat Pratupa community have defaulted towards more open sharing of heritage documentation, indicative of diverse approaches to safeguarding in the digital age.

In this paper, therefore, I look to a range of participatory media projects, including those I have described above, to argue that community-based productions of multimedia aimed at documenting, transmitting, and revitalizing intangible heritage are significant sites in which local cultural property discourses are articulated and put into practice. This is particularly important in the age of the 'born digital' ethnographic object, where heritage documentation can become subjected to unlimited circulation in the form of digital copies and remixes. National and local governments, heritage workers, anthropologists, curators—and, increasingly local stakeholders who represent their own cultures, languages, and histories—are some of the agents of transformation of intangible cultural expression into digital heritage. All play a role in determining what media documentation enters the public domain, and what remains privately managed at the local level. As Dorothy Noyes has argued, rather than reifying the concept of tradition as community managed heritage, folklorists, anthropologists, heritage workers, and policy makers should instead view local tradition as a vehicle for the "collective negotiation of

⁴ Doig River First Nation, Amber Ridington, and Kate Hennessy, *Dane Wajich—Dane-zaa Stories and Songs: Dreamers and the Land*. (Virtual Museum of Canada, 2007).

<http://www.museevirtuel-virtualmuseum.ca/sgc-cms/expositions-exhibitions/danewajich/english/index.html>

⁵ For detailed ethnographies of Aboriginal prophet movements in the Canadian subarctic, see: Robin Ridington, *Trail to Heaven: Knowledge and Narrative in a Northern Native Community* (Vancouver and Toronto: Douglas and McIntyre, 1988); Jean-Guy Goulet, *Ways of Knowing: Experience, Knowledge, and Power Among the Dene Tha*. (Lincoln: University of Nebraska Press, 1998); June Helm, *Prophecy and Power Among the Dogrib Indians* (Lincoln: University of Nebraska Press, 1994).

⁶ For more information and e-learning resources from the Princess Maha Chakri Sirindhorn's Intangible Heritage and Museums Field School, please visit <http://www.sac.or.th/databases/fieldschool/>

intracommunity conflict.”⁷ The projects I describe highlight processes of intra and inter-community negotiation of sensitive issues of ownership and cultural knowledge; while the digital is a common source of tension and anxiety, local responses and decisions vary across cultural, geographical, and historical contexts.

World heritage policies represent another dimension of the conversation. The UNESCO *Charter on the Preservation of Digital Heritage*, for example, states that “access to digital heritage materials, especially in the public domain, should be free of unreasonable restrictions. At the same time, sensitive and personal information should be protected from any form of intrusion.”⁸ But how, and when, is documentation of sensitive digital heritage differentiated from that suitable to be circulated in the public domain? In the name of safeguarding cultural heritage, how are decisions made about what should be made public, and which should be kept private? How can emerging anxieties and conflicts be productively channeled into the articulation of local cultural property rights discourse aimed at safeguarding? I argue that participatory media production projects aimed at documenting and transmitting cultural heritage create opportunities for this kind of decision making. The 2003 UNESCO *Convention for the Safeguarding of the Intangible Cultural Heritage*, with its emphasis on the role of local communities as full partners in efforts to safeguard their cultural heritage,⁹ must be considered in relation to the universal access oriented discourse around digital cultural heritage. In the age of the ‘born digital’ ethnographic object, the safeguarding of the intangible and the digital must be understood as interwoven projects.

2. The Intangible and the Digital

While documentation of intangible heritage is only one aspect of safeguarding, it represents an important moment in the transition from intangible expression to digital cultural heritage. The proliferation of digital tools available at low cost for the increasingly interconnected projects of documentary recording, archiving, and sharing has amplified the scale of digital production in heritage conservation initiatives¹⁰ and has implicated digital documentation in processes of making media public and removing control over the circulation of heritage from local contexts.

In the following section, I look to UNESCO’s definitions of *intangible cultural heritage*, *safeguarding*, and *digital heritage* to emphasize the role of participatory media production projects in creating space in which key decisions can be made about the ethical circulation of heritage documentation. How do these heritage policy definitions relate to a spectrum of on-the-ground practices of knowledge and information management in diverse global contexts? I begin to answer this question by describing a range of participatory digital media-based projects that offer insight into the potential of digital production to further local goals for ‘safeguarding’ cultural heritage in the digital age.

⁷ Dorothy Noyes, “The Judgment of Solomon: Global Protections for Tradition and the Problem of Community Ownership,” *Cultural Analysis* 5 (2006): 28.

⁸ UNESCO, *Charter on the Preservation of the Digital Heritage* (2003), Article 2.

⁹ See Richard Kurin, “Safeguarding Intangible Cultural Heritage: Key Factors in Implementing the 2003 Convention,” *International Journal of Intangible Heritage* 2 (2007): 10-20.

¹⁰ A spectrum of projects, approaches, and critical responses to the use of new technologies in heritage conservation can be found in: *Theorizing Digital Heritage: A Critical Discourse*, ed. Fiona Cameron and Sarah Kenderdine (Cambridge and London: MIT Press, 2007); *New Heritage: New Media and Cultural Heritage*, ed. Yehuda E. Kalay et al. (London and New York: Routledge, 2008).

“Intangible cultural heritage” is defined in the 2003 Convention for the Safeguarding of the Intangible Cultural Heritage as:

...the practices, representations, expressions, knowledge, skills – as well as the instruments, objects, artifacts and cultural spaces associated therewith – that communities, groups and, in some cases, individuals recognize as part of their cultural heritage. This intangible cultural heritage, transmitted from generation to generation, is constantly recreated by communities and groups in response to their environment, their interaction with nature and their history, and provides them with a sense of identity and continuity, thus promoting respect for cultural diversity and human creativity.¹¹

The Convention also details a process through which intangible heritage may be protected for future generations. The notion of ‘safeguarding’ in the Convention is described as:

...measures aimed at ensuring the viability of intangible cultural heritage, including the identification, documentation, research, preservation, protection, promotion, enhancement, transmission, particularly through formal and non-formal education, as well as the revitalization of various aspects of such heritage.¹²

Article 15 of the Convention further emphasizes the role of cultural communities, groups and individuals in safeguarding initiatives, stating that:

Within the framework of its safeguarding activities of the intangible cultural heritage, each State Party shall endeavor to ensure the widest possible participation of communities, groups, and where appropriate, individuals that create, maintain, and transmit such heritage, and to involve them actively in its management.¹³

Safeguarding, by these definitions, should depend on participation at the local level, rather than top-down intervention and control of intangible heritage initiatives. As Richard Kurin has emphasized, this Article in the Convention represents a shift in perspective on the role of culture bearers in determining best practices for safeguarding. He writes:

Governments, or university departments or museums, cannot just assume that they have permission to define intangible cultural heritage and undertake its documentation, presentation, protection, or preservation. Community participation is meant to be significant and meaningful—involving the consent of community leaders, consultation with lead cultural practitioners, shared decision making on strategies and tactics of safeguarding and so on. Article 15 strongly empowers the community in the operation of and realization of the Convention.¹⁴

Local participation in safeguarding initiatives must therefore include more than decision making about what to include in inventories and lists of intangible heritage, or what to document; indeed, as Michael Brown points out, the discipline of anthropology “long ago concluded that documentation has only a

¹¹ UNESCO, *Convention for the Safeguarding of the Intangible Heritage*, 2.

¹² UNESCO, *Convention for the Safeguarding of the Intangible Heritage*, 9.

¹³ UNESCO, *Convention for the Safeguarding of the Intangible Heritage*, Article 15.

¹⁴ Richard Kurin, “Safeguarding Intangible Cultural Heritage: Key Factors in Implementing the 2003 Convention,” *International Journal of Intangible Heritage* 2 (2007), 15.

modest role in the preservation of culture. To think otherwise is to make the classic error of mistaking the map for the territory it represents.”¹⁵ Rather, participatory processes of safeguarding should involve the creation of opportunity for the careful consideration of the implications of digital documentation, and the development of local strategies for determining which documentation can safely enter the public domain.

These considerations are important as the “digital heritage” becomes a focus of international preservation efforts. Digital heritage is defined in the draft Charter on the Preservation of the Digital Heritage as consisting of:

...unique resources of human knowledge and expression. It embraces cultural, educational, scientific and administrative resources, as well as technical, legal, medical and other kinds of information created digitally, or converted into digital form from existing analogue resources. Where resources are “born digital”, there is no other format but the digital object.

Digital materials include texts, databases, still and moving images, audio, graphics, software and web pages, among a wide and growing range of formats. They are frequently ephemeral, and require purposeful production, maintenance and management to be retained.

Many of these resources have lasting value and significance, and therefore constitute a heritage that should be protected and preserved for current and future generations. This ever-growing heritage may exist in any language, in any part of the world, and in any area of human knowledge or expression.¹⁶

Like the definition of intangible cultural heritage, which seems to embody nearly every possible form of expression, so the digital heritage would seem to include nearly all of contemporary digital production. “Born digital” media—which represents an exponentially growing domain of digitally produced documentation of intangible cultural heritage—fits neatly into this definition. However, the Digital Heritage Charter also acknowledges the complexities of legal and ethical access to digital materials:

The purpose of preserving the digital heritage is to ensure that it remains accessible to the public. Accordingly, access to digital heritage materials, especially those in the public domain, should be free of unreasonable restrictions. At the same time, sensitive and personal information should be protected from any form of intrusion.

Member States may wish to cooperate with relevant organizations and institutions in encouraging a legal and practical environment which will maximize accessibility of the digital heritage. A fair balance between the legitimate rights of creators and other rights holders and the interests of the public to access digital heritage materials should be reaffirmed and promoted, in accordance with international norms and agreements.¹⁷

While the intricacies of implementing and policing access to digital heritage materials remain to be explored, I hold these definitions up against one another to emphasize the extent to which the intangible

¹⁵ Michael Brown, “Heritage Trouble: Recent Work on the Protection of Cultural Property,” *International Journal of Cultural Property* 12 (2005), 48.

¹⁶ UNESCO, *Charter on the Preservation of the Digital Heritage*, 2003, 1.

¹⁷ *Ibid.*, 2.

and the digital, and their related policy instruments and definitions, are connected through the act of documentation and circulation. The complexities of safeguarding intangible cultural heritage in the digital age are not adequately reflected in current policy documents, even though the need for stakeholder participation and attention to digital heritage access are acknowledged. Museum scholar Fiona Cameron laments the lack of critical discourse around digital heritage, even though the “ascription of heritage metaphors to cultural materials in a digital format means that digital media has become embedded in a cycle of heritage value and consumption, and in the broader heritage complex.”¹⁸ With emphases on ensuring maximum public access through programs like “Information for All” and “Memory of the World”, the UNESCO *Charter on the Preservation of the Digital Heritage*, for example, is seen to exemplify the induction of digital cultural heritage materials into broader dynamics of globalization and heterogenization.¹⁹ As Michael Brown points out, major heritage policy documents demonstrate a tension between cultural internationalists and cultural nationalists, an ongoing concern with “the balance between heritage as a resource for all of humanity and something that properly belongs to, and remains controlled by, its community of origin.”²⁰ Jane Anderson has further explored the ‘anxieties’ associated with access to and circulation of the contents of colonial archives; she describes a growing tension in which Indigenous communities are demanding recognition as legitimate authors and owners of documents representing their cultures, but are faced with the reality that legal ownership is granted to the individual who made the documentation (a photograph, an audio recording, a video recording, and so on). According to Anderson, these archival materials are anxiety inducing because they often do not reflect contemporary cultural identifications and desired representation, or their anticipated use and circulation.²¹ Participatory processes can represent moments of negotiation—even conflict—over what to circulate publicly and what to manage privately, determining how and if the products of intangible cultural heritage safeguarding initiatives should become a part of the world’s digital heritage. The resulting tensions and anxieties are exacerbated in discourse and practices related to the production of digital heritage, making local participation in documentation of the intangible and the digital increasingly relevant.

3. Participatory Media Production and Local Cultural Property Rights Discourse

As I will describe below, documentary practices and related negotiations, conflicts, and dynamics create opportunity for the discussion of ownership and ethical circulation of cultural property. The following case studies represent a spectrum of techno-mediated approaches to safeguarding heritage in the digital age, in which the articulation of local cultural property rights discourse plays a central role. I begin with a series of examples from the Pacific, North America, and Australia, and then move on to describe my own fieldwork in northern British Columbia and Thailand.

¹⁸ Fiona Cameron, “The Politics of Heritage Authorship: The Case of Digital Heritage Collections” in *New Heritage: New Media and Cultural Heritage*, ed. Yehuda E. Kalay et al. (Cambridge and London: MIT Press, 2007), 71.

¹⁹ Fiona Cameron, “Beyond the Cult of the Replicant: Museums and Historical Digital Objects—Traditional Concerns, New Discourses,” in *Theorizing Digital Cultural Heritage: A Critical Discourse*, ed. Fiona Cameron and Sarah Kenderdine (Cambridge and London: MIT Press, 2007).

²⁰ Michael Brown, “Heritage Trouble: Recent Work on the Protection of Cultural Property,” *International Journal of Cultural Property* 12 (2005), 48.

²¹ Jane Anderson, “Anxieties of Authorship in the Colonial Archive,” in *Media Authorship*, ed. C. Chris and D. Gerstner (London: Routledge, forthcoming 2012).

As a first example, Guido Pigliasco's collaborative production of a cultural heritage DVD and archive with the Sawau Tribe of the island of Beqa, Fiji became a process of negotiation of intellectual property issues.²² *The Sawau Project* focuses on the reclamation of documentation of the *vilavilavevo*, the Sawau practice of firewalking. In past decades, the *vilavilavevo* has been widely circulated and commodified, but has now been claimed by the tribe as their own to control and perform.²³ The DVD project, which uses the geography of Beqa as its framework for navigation of content, is an archive of repatriated documentation of the *vilavilavevo* and video vignettes detailing the origins of the firewalking tradition. Engaging with this media and the meaning of this element of Sawau intangible heritage required the negotiation of anxieties associated with the sharing of digital media. The successful completion of the DVD project required the collective expression of a local cultural property rights discourse to make decisions about what could be shared, and what would be safeguarded offline. Ultimately, project participants made the decision to limit the circulation of their project, restricting it to locally shared DVDs instead of web-based access, thereby reducing participation in the ongoing appropriation of practices belonging to the Sawau people.

In another example, Jason Baird Jackson describes Woodland Indian digital documentary practices in the contexts of cultural performance and ritual.²⁴ He makes the observation that as new recording technologies have become available, Indian peoples in Oklahoma who are concerned with the conservation of ancestral forms of music, dance, and ritual have actively integrated digital documentation into their production of digital archives for educational and personal use. He notes that these practices have emerged along with anxiety and tension about the potential commoditization of documentation, and the loss of ceremonial leaders to control how recordings are used. Woodland digital documentary practices, which include the use of cell phones, video cameras and other ubiquitous technologies "...have unfolded within a local intellectual property (IP) system rooted more broadly in tribal regional cultures and social norms."²⁵ At the same time, Jackson argues that these same practices can be found to be in contravention of World Intellectual Property Organization (WIPO) treaties, therefore potentially subject to massive, bankrupting fines—demonstrating significant tensions between local practices and regimes of ownership, and global heritage policies that aim to protect these same practices. Safeguarding cultural heritage, in these contexts, is more complex than ascribing to specific international policies; as Jackson convincingly argues, both the Free Culture movement (as described by Lawrence Lessig, for example²⁶) and intellectual property solutions presented by WIPO and others place local communities in contradictory positions that have yet to find resolution.

Kimberly Christen's work with Warumungu people in Australia, and with Native American tribes in the United States, further shows how the collaborative design and implementation of digital heritage archives can create opportunities for the negotiation and articulation of local cultural property rights

²² Guido Carlo Pigliasco, "Intangible Cultural Property, Tangible Databases, Visible Debates: The Sawau Project," *International Journal of Cultural Property* 16 (2009):255-272.

²³ Kate Hennessy, "A Ituvatuva Ni Vakadidike E Sawau: The Sawau Project DVD," *Visual Anthropology Review* 25(1, 2009): 90-92.

²⁴ Jason Baird Jackson, "Boasian Ethnography and Contemporary Intellectual Property Debates," *Proceedings of the American Philosophical Society* 154 (1, 2010): 40-49.

²⁵ *Ibid.*, 44.

²⁶ Lawrence Lessig, *Remix: Making Art and Commerce Thrive in the Hybrid Economy* (New York: The Penguin Press, 2008).

discourse.²⁷ Her work on the development of the *Mukurtu*²⁸ open-source cultural heritage management system has generated insight into the wide spectrum of approaches to local heritage management that Indigenous peoples in particular are seeking to meet their contemporary needs for safeguarding their cultural property. Mukurtu gives local communities the opportunity to define culturally appropriate access to heritage documentation by customizing the Mukurtu database to meet their particular needs. In this way users are able to engage in decision making around the definition what should be made public, and what should remain private, or which media should remain somewhere in between, circulating within the “proper” systems of knowledge exchange and supporting off-line, everyday social and cultural interactions that involve the limited exchange of cultural knowledge. The Mukurtu archive, and the related Plateau Peoples’ Web Portal are both built on principles of “respectful repatriation” that aim to support such ethical circulation of cultural property.²⁹

3.1 From Canada to Thailand

These dynamics of media production and negotiation of representation of culture and language have also been reflected in my fieldwork in Canada and Thailand. For example, linguistic anthropologist Patrick Moore and I identified similar outcomes in our study of endangered language documentation among members of the Carcross-Tagish First Nation in the Yukon Territory in northern Canada.³⁰ Working in partnership with the First Peoples’ Cultural Foundation to document the Tagish language, and upload documentation to the language archiving website FirstVoices.com, we observed that project participants created an environment in which elders and youth were able to articulate an Indigenous language ideology in resistance to the values and historical practices of residential schools and a history of control of native language revitalization by outside organizations. In this environment, local control over language revitalization efforts was facilitated by a collaborative relationship with archiving and technical consultants at FirstVoices.com. Participants placed emphasis on the holistic nature of language and culture, showed preference for traditional modes of social interaction, and demonstrated the centrality of elders’ knowledge in the language revitalization process. The digitally mediated space created through partnership with the First Peoples’ Cultural Foundation also functioned to connect language revitalization efforts to broader Carcross-Tagish discourse around political authority, land claims, and cultural identity.

During my work with the Doig River First Nation between 2004 and 2008, I co-curated (with Amber Ridington) a collaboratively produced Virtual Museum of Canada exhibit of oral narratives and song called *Dane Wajich—Dane-zaa Stories and Songs: Dreamers and the Land*.³¹ The project drew on archival ethnographic documentation created in Dane-zaa communities by anthropologists Robin Ridington, Jillian Ridington, and Antonia Mills over the last forty years.³² It re-presented these media

²⁷ Kimberly Christen, “Opening Archives: Respectful Repatriation,” *American Archivist* 74 (2011): 185-210.

²⁸ *Mukurtu CMS*, <http://www.mukurtu.org/>. Accessed Aug. 24, 2012.

²⁹ Kimberly Christen, “Opening Archives: Respectful Repatriation,” *American Archivist* 74 (2011): 185-210.

³⁰ Patrick Moore and Kate Hennessy, “New Technologies and Contested Ideologies: The Tagish FirstVoices Project,” *American Indian Quarterly* 30(1&2) (2006):119-137.

³¹ Doig River First Nation, Amber Ridington, and Kate Hennessy, *Dane Wajich—Dane-zaa Stories and Songs: Dreamers and the Land*. (Virtual Museum of Canada, 2007).

<http://www.museevirtuel-virtualmuseum.ca/sgc-cms/expositions-exhibitions/danewajich/english/index.html>

³² Robin Ridington and Jillian Ridington, “Archiving Actualities: Sharing Authority with Dane-Zaa First Nations,” *Comma* 1 (2003):61-68; Robin Ridington and Jillian Ridington, *When You Sing it Now, Just Like New: First Nations Poetics, Voices, and Representations* (Lincoln and London: University of Nebraska Press, 2006).

alongside contemporary documentation of narrative and song that we had created in collaboration with members of the Doig River community in the course of the virtual exhibit production.

As I have described elsewhere,³³ *Dane Wajich* was developed using an open, participatory production process that was guided by elders and community leaders. It involved Doig River First Nation youth in central roles as media documentarians and organizers. Project planning meetings were recorded, some of which became central elements of the virtual exhibit for the way they demonstrated a Dane-zaa methodology for intangible heritage documentation. One such meeting took place in July of 2005 in the Doig River gymnasium. Elder Tommy Attachie spoke to a group of community members and project participants assembled to plan the project. His focus was a painted moose hide drum skin that had been made by a Dane-zaa prophet named *Gaayeq* one hundred years before; the drum had been brought to a meeting by its caretaker, former Chief Garry Oker, who played a central role in project planning. Tommy Attachie used the drum, which featured a map of heaven dreamed and painted by *Gaayeq*, to connect the history of Dane-zaa prophets to material culture, oral narrative, and land. He also used his knowledge of the drum—its history, and its significance in the present—to define a methodology for documenting Dane-zaa intangible cultural heritage. In the weeks that followed the performance of this narrative, our group traveled to seven sites in Dane-zaa territory where elders, youth, ethnographers and linguists recorded videos documenting life histories, traditional narratives, and histories of Dane-zaa dreamers. Between 2005 and 2007, our project team then worked to develop and complete the virtual exhibit, consulting with Chief and Council, elders, and community members from the storyboarding pre-production process through to official exhibit launch.

In the course of these consultations and design sessions, however, important questions were raised about the ownership and control of archival recordings of Dane-zaa dreamers, as well as the appropriateness of showing images of dreamers drawings, like the one featured on *Gaayeq*'s drum, to the public over the Internet. Objections were raised in a neighboring Dane-zaa community by descendants of another prominent dreamer, Charlie Yahey, about the use of the image of their ancestor in Doig River's media project. Ultimately, the Doig River Chief and Council had to make a decision, taking into account varying positions held by members of the community, about how to proceed. The decision was made to remove all images of dreamers drums from the virtual exhibit, in order to respect traditional care and handling of dreamers drums, which had included keeping them out of public view, except in special circumstances. The decision was also made to respect the intellectual property rights being claimed by the descendants of Charlie Yahey, which meant no longer using media that documented him in the exhibit. At Doig River, meaningful local participation in digital documentation of intangible cultural heritage, and the subsequent presentation of oral narratives, photographs, and other media, opened up space for negotiation and debate over ownership and control of Dane-zaa cultural heritage and its circulation in digital form.

I observed similar dynamics in my fieldwork in Thailand. Between 2009 and 2011, I worked as a lecturer and resource person in the Intangible Cultural Heritage and Museums Field School in Lamphun, northern Thailand, organized by the Sirindhorn Anthropology Center and UNESCO, Bangkok. The goal

³³ Kate Hennessy, "From Intangible Expression to Digital Cultural Heritage," in *Safeguarding Intangible Heritage*, ed. Michelle Stefano et al. (London: Boydell and Brewer, 2012), 33-46; Amber Ridington and Kate Hennessy, "Building Indigenous Agency through Web-based Exhibition: Dane Wajich—Dane-zaa Stories and Songs: Dreamers and the Land," *Proceedings of Museums and the Web*, 2008. Accessed Aug. 22, 2012. <http://www.museumsandtheweb.com/mw2008/papers/ridington/ridington.html>

of the fieldschool is to introduce students from the Mekong Delta region to a wide range of practical issues, debates, case studies, and critiques of the 2003 ICH Convention. Over the last two field school seasons, I collaborated with students and the temple community of Wat Pratupa, one of the field school research sites, to document and represent locally-identified endangered elements of intangible heritage. At Wat Pratupa, local heritage documentation activities and digital circulation of representations of intangible heritage over the Internet have created opportunities for the negotiation and articulation of local cultural property rights discourse.

Wat Pratupa's Assistant Abbott, Phra Patiphan Puriphanyo, is the creator and Webmaster of a site called www.muanglamphun.com, on which, at that time, he regularly posted local documentation and news related to temple activities and documentation of local traditions, festivals, and practices. The website, and its related Facebook page, were being used as a strategy for circulating the distinct practices of Wat Pratupa's ethnic Yong community. In 2010 and 2011, I worked with students to explore some of these practices; first, we looked at issues related to the safeguarding of ethnic Yong poetic narratives called *Kap Kalong*, which reiterate the history of ethnic Yong migration from Burma, and details contributions of families and individuals to the Buddhist merit-making festival, the *Salak Yom*; the following year, we collaborated with the Assistant Abbott and community members to produce a short documentary video about local efforts to revitalize and protect the endangered Yong language.³⁴ This video and other field school documentation were circulated on www.muanglamphun.com and the related Facebook page.

Wat Pratupa's current approach to sharing documentation of their intangible heritage, I learned, was largely shaped in the process of developing the temple website. The site and Facebook page first featured extensive photographic documentation of the *Salak Yom* festival, historical photographs that the Assistant Abbott had collected from members of the community, and representations of other local traditions that had been identified as in need of protecting and publicizing. However, it was the decision to document and share images of sacred material culture owned by Wat Pratupa—specifically, the contents of a Buddhist palm-leaf manuscript and a rare wooden Buddha carving—that stimulated local discussion about the benefits and risks of digital documentation in the service of safeguarding heritage. After images of the Buddha and manuscript were posted on the website, villagers were surprised by the outside interest they generated, including the arrival of non-local filmmakers seeking to make a documentary about the valuable collection. With new awareness of the digital visibility of the collection, members of the community began to worry about the physical safety of the objects. Yet these events, and the anxieties that they produced, resulted in local decision making about appropriate digital circulation of heritage documentation. Eventually, the Assistant Abbott told me, it was decided that it was advantageous to share images of significant sacred objects in the Wat Pratupa collection, because the original objects could be kept safe and out of public view, protected from thievery. He told me:

There's an idiom saying, 'If you swallow, it disappears; if you spit it out, it remains'. No matter how wise you are, if you don't have a disciple, your wisdom goes to waste. But if you teach your disciples, your knowledge transcends your own life. It doesn't matter if replicas were created, because a genuine is still a genuine. In fact, there are even more watchers than before because there are more people who are aware of these significances.

³⁴ Sirindhorn Anthropology Centre, "Because We Are Yong," Video, 2011 (7 mins 12 sec). Produced by Tashi Dendup, Kate Hennessy, Nalina Gopal, Aynur Kadir, Inpone Thephetlasy, and Aree Tirasatayapitak, and Apinan Thammasena. Accessed Aug 20, 2012, at <http://www.youtube.com/watch?v=jGwgLiJ1bYE&feature=plcp>

A sense of ownership keeps growing, which may lead to two different strategies: increased security measures, or increased studies and revitalization. The decision depends on the conservators and the community... Let the knowledge spread in the community.³⁵

4. Conclusion

In this paper, I have presented examples from the field that locate the production and circulation of digital heritage as central in debates over local cultural property rights. Participatory media production processes, where the transformation of intangible cultural heritage into digital heritage takes place, represent important moments in which community-based negotiation of the control of documentation and cultural representation can take place. These negotiations, as I have described, can include the definition of appropriate methodologies for intangible heritage documentation and the digital circulation of representations of local material culture, archival media, and intangible expressions. These locally defined processes of cultural heritage documentation and their negotiations facilitate the development of local approaches to controlling cultural property, which will range from restrictive to more liberal, depending on content and context. These processes and negotiations are particularly important in relation to world heritage policies, leading to the question: how can international policy instruments better acknowledge and support the range of on-the-ground approaches to cultural property and safeguarding? The Intangible Cultural Heritage Convention and the Digital Heritage Charter should be considered and implemented with awareness of the complexities and diversity of local cultural property rights discourse. State parties, heritage workers, and community members should be encouraged to take necessary steps to ensure that meaningful participation in intangible and digital heritage safeguarding initiatives includes space for the negotiation of diverse approaches to ownership and circulation of cultural heritage.

Acknowledgements

Sincere thanks to members of the Doig River First Nation, to *Dane Wajich* exhibit co-curator Amber Ridington, and to Patrick Moore, Robin Ridington, and Jillian Ridington. Thank you to Wat Pratupa Assistant Abbott Phra Patiphan Puriphanyo. My gratitude to the organizers of the Intangible Heritage and Museums Field School in Lamphun, Thailand, and the staff of the Princess Maha Chakri Sirindhorn Anthropology Centre in Bangkok, particularly Alexandra Denes, Suvanna Kriengkraipetch, Paritta Chelernpow Koanantakool, and Linina Phuttitarn.

³⁵ Phra Patiphan Phangwana, Interview conducted by Kate Hennessy, August 18, 2011, Lamphun, Thailand. Translation by Linina Phuttitarn.

Preservation Infrastructures: Current Models and Potential Alternatives

An Example to Follow

An Infrastructure for Interoperability and Governance in the Tuscan Public System for Digital Preservation

Ilaria Pescini and Walter Volpi

Regione Toscana, Italy

Abstract

For years, Tuscany has promoted several projects of digitization of administrative procedures. These projects are based on a technological and organizational pre-existing substrate: a territorial system that involves government agencies of that region and private. Entire community uses that same technological infrastructures that are shared across regional territory and which allows the creation of cooperative services using normalized and standardized rules and languages. On this basis the project DAX (Digital Archives EXtendend) was started. This project has led to the creation of an infrastructure for long-term preservation for digital archives. It serves all the regional administration of Tuscan territory. DAX is an accountable system to describe and manage non-current and historical archives, and to storage singles records. DAX is an example of cooperation and interoperability policies pursued for years in Tuscany, in the field of innovation.

Authors

For the past ten years Ilaria Pescini has been an employee of the Regional Government of Tuscany, where she has the responsibility for archives and document management system. She had already a long experience with traditional historical archives. She has been increasingly involved in digitization, a process triggered by national laws issued since 2000. She is now focusing on projects that relate to record management, digital document management (authenticity, electronic signature, interoperability, workflow management), digital certificates, and citizen-oriented services for the free but secure circulation of documents. She has collaborated closely with the ICT division of the regional government. She is member of regional and national working group about archival standard and archival legislation.

Walter Volpi has a degree in Computer Science from the University of Pisa; for the past ten years he has been an employee of the Regional Government of Tuscany, where he is responsible of interoperability and applicative cooperation unit. He is a supervisor of IT architecture, focused on event driven architecture (EDA) and service oriented architecture (SOA) and a Service Level manager for the territorial infrastructure of interoperability and applicative cooperation. He leads regional and national working groups with representatives of the main ICT vendors, universities, central and local administrations, in order to define rules and models for interoperability standards both technical and semantical. He is Project Manager that relate to digital document management, digital preservation and Open data.

1. Introduction

This report provides an overview about the system for long-term preservation of digital archives¹ by the Region of Tuscany. The system will storage the archives by all different government agencies within the region. This is one of the first experiences in this branch in Italy and it is a meaningful experience for at least four reasons, generally speaking and not specific about digital preservation.

¹ In this report we intend to speak about archive like a whole of the records organically created and accumulated by a public corporate body in the course of that creator's administrative activities.

The first reason concerns organization and descends from the choice of creating a territorial system involving all the various government agencies of the region, under the coordination of the main administration. The second reason is based on political choices that encourage the development of public governments network sharing many technological innovation projects. The third reason, of technical nature, provides a valuable contribution to the entire system: the technological infrastructures the system uses, work for the entire community. These infrastructures implement the interoperability and interchange channels for many other services.

Furthermore, as fourth reason, it obeys the Italian legislation on digitization² and, at the same time, interprets the State law which delegates each public administration to maintain their own archives, introducing a new perspective where a single regional archive assembles them.

2. The Context

It is important to underline that the long-term preservation system, a project created, commissioned and planned by Region of Tuscany and called DAX (Digital Archives eXtended), was conceived in an advanced administrative, archival and technological context.

Regione Toscana is a territorial administration created in 1970, like all the other regions located all over the Italian national territory, and it is one of the five kind of institutions constituting Italian Republic. The task of each Region is the government and the growth of its territory; this is possible thanks to its strong legislative power and to an administrative and planning organization, which are both exercised in complete independence from the central power, but with a strong link with local authorities. In comparison with other Italian Regions, Tuscany made a stronger use of tight cooperation with local institutions³ to exercise its governance through agreements on many different topics. Regione Toscana has been the main driving entity for its territory concerning the themes of innovation mainly through the Information and Communication Technology (ICT) project. For this reason, Regione Toscana is a national model and a reference point regarding e-government.⁴

Since 2004 Regione Toscana, through a regional law,⁵ has established a community network that gather up all the public institutions forming the regional territory, i.e., local administrations, Tuscan universities and public bodies dealing with public healthcare. Following this law the Community Network

² The term dematerialization indicates the gradual increase in the computerized document management - within public and private administrative structures - and the consequent replacement of traditional media in favor of electronic documents. Dematerialization is one of the central topic of the Italian legislation on reform and innovation in public administration (see Codice Amministrazione Digitale – Dlgs March 7, 2005, n. 82). Since 1997 the Italian national legislation recognizes full legal value to electronic documents.

³ In the Italian legal system a local authority (ente locale) is a public body whose jurisdiction is limited within a specific territorial area. The local authorities are opposed national bodies whose competence extends over the entire national territory. In Italy, the term has a specific meaning referring to local authorities such as municipalities, provinces and metropolitan cities, under the Italian Constitution.

⁴ E-Government (short for electronic government) is digital interactions between Governments or Agencies, and between government and the Citizens or businesses. It is defined as “The employment of the Internet and the world-wide-web for delivering government information and services to the Citizens.” (United Nations Department of Economic and Social Affairs, “United Nations E-Government Survey 2012”).

⁵ Regional Law of Tuscany Region 26 January 2004, No. 1, “The promotion of electronic administration and of the information and knowledge society throughout the regional system. Rules for the “Tuscany Region Data Communication Network.”

has its own managing structure, and promotes cooperation among different administrative entities, according to his own skills and necessities. Operational tools and functioning bodies of this network involve politicians and technicians of all the administrations: a Strategic Committee, which plans and coordinates the program and subjects, drives the choices implemented by a Technical Direction through the promotion and the realization of projects and services for every joining member; at last a general Assembly gathers all the administrations and, once a year, introduces and evaluates the results. The Network gives legal identification to systems of administrative interoperability, of instruments and contents sharing, of data and information and of administrative processes.

The kind of relations created among public institutions and different subjects⁶ form a complex system, managed through ICT systems and architectures.

The computerization process started to yield significant results thanks to the new transversality and interoperability of technologies, which give concreteness to the interoperability and transversality of administrative effort, to which we are referring.

Italian national rules, concerning public administrations technologies, have been promoting and stimulating the use of interoperable systems and infrastructures for years; the aim of these technological infrastructures is the integration of procedures used by different subjects of the same territory (both national or regional). Through these technological infrastructures is possible to achieve data and information interchange and interaction among different entities allowing them to cooperate.⁷ In Tuscany this is supported by a unique technological infrastructure that carries out applicative cooperation assuring an extremely advanced interoperability.⁸

Every dematerialized administrative process becomes part of a meta-system that allows the sharing of tools and information, coming from different administration systems included in the same Community. This technological choice comes from an organizational and archival need, because information, documents and administrative processes have a transversal role among more entities that are often part of the process, at the same responsibility level.

Tuscan system perfectly responds to the recommendations of European Interoperability Framework for Pa European e-Government Services (EIF)⁹ which identifies different levels of interoperability and emphasizes the importance of cooperation among organizations and processes coordination.

⁶ Since the administrative reforms of the late nineties has been introduced, in the functioning of the Italian public administration system, a kind of institutional relations network, which reconciles the need for autonomy and accountability with the need for integration based essentially on the system of local government and in accordance with the principle of subsidiarity. The Italian public administration today is seen as a unique network of subjects that intersect powers and functions, subjects who are no longer part of a hierarchy but are coordinated in a network system in which each element is a sibling node.

⁷ <http://www.progettoicar.it/Home.aspx>

⁸ <http://www.cart.rete.toscana.it>

⁹ For European Interoperability Framework the dimension of interoperability are: *Political Context* - Cooperating partners having compatible visions, and focusing on the same things; *Legal Interoperability* - The appropriate synchronization of the legislation in the cooperating MS so that electronic data originating in any given MS is accorded to proper legal weight and recognition wherever it needs to be used in other MS; *Organisational Interoperability* - The processes by which different organisations such as different public administrations collaborate to achieve their mutually beneficial, mutually agreed eGovernment service-related goals; *Semantic Interoperability* - Ensuring that the precise meaning of exchanged information (concept, organisation, services, etc.) is preserved and well understood; *Technical Interoperability* - The technical issues involved in linking computer systems and services (open interfaces, interconnection services, data integration, middleware, data presentation and exchange, accessibility and security services, ...). Cfr. <http://ec.europa.eu/idabc/en/chapter/5883.html>.

On archival and normative side, Region of Tuscany declared, with regional law of 2009, to “take necessary measures for the de-materialisation of administrative documents, encouraging their storage in digital format with methods which enable preservation and use over time.”¹⁰ Moreover Regione Toscana “provides for and maintains a technological platform and digital services for the preservation of computer documents which enables joint management of the documents in both hard copy and digital format...”¹¹ With the same law, and consistently with what established by the territorial governance, the Region “promotes the establishment of the regional administration and regional agencies archive network in order to favour the sharing of tools and information in a coordinated manner, as well as access to the archive documentation and the development of documentary assets” and the improvement of documental heritage.¹²

Through DAX Region of Tuscany has materialized regional law 2009, n. 54.

3. Motivations and Goals

Within this organizational and technological context, the regional Community Network has delegated the Region of Tuscany to build a platform for the preservation of administrative records¹³ produced in digital form by Tuscan public administrations. So, DAX arose in response to the needs of a variety of subjects, even if the major administration—the Region—has played a larger role in coordinating the activities: it has the responsibilities in the design and implementation of the system. In particular the Region takes care of the dissemination of the culture of these issues. It should not be forgotten that such a complex system would be difficult to achieve for small administrations as Italian municipalities are. A coordinated project helps the cheapness and a considerable saving in terms of human resources and management.

It seemed also important for the growth of the area and its public administration, in this moment of transition from traditional to digital documents, to develop the culture of these issues and also provide support to smaller organizations, creating a common cultural fabric and shared rules. In addition to the uniformity of treatment of digital archives, a single storage system would have reached a higher level of performance in the efficiency of public administration and most of all, the easier relations with citizens, and with all users interfacing with a unique system.¹⁴

¹⁰ Regional Law of Tuscany Region, 5 October 2009, n. 54, “Establishment of the regional information and statistical systems. Measures for the coordination of infrastructures and services for the development of the information and knowledge society,” art. 10 (“Documentary activities”), comma 2.

¹¹ Regional Law of Tuscany Region, 5 October 2009, n. 54, ..., art. 10 (“Documentary activities”), comma 3.

¹² Regional Law of Tuscany Region, 5 October 2009, n. 54, ..., art. 14 (“Regional Archives”).

¹³ With the expression “administrative document” we intend to indicate any graphic, fotocinematografic, electromagnetic or any other species of the content of documents, including internal or not related to a specific process documents. They are held by a public authority and related public interest activities.

¹⁴ “Access right” means, in accordance with current Italian legislation, the right of interested parties to consult and take copies of administrative documents. All citizens, companies and associations, including those of public or common carriers, can exercise access right (see http://www.governo.it/Presidenza/DICA/4_ACCESSO/). The access right to administrative documents is a right granted to citizens on the basis of relations with the state and public administration, in order to ensure transparency of the governments. In Italy the access right is enshrined in Italian law, 7 August 1990, n. 241 “New Rules Regarding Administrative Procedure and the Right of Access to Administrative Documents.”

We cannot forget that archive has an important role in identity and memory of the activities of its producer, and that the necessity to preserve documents has to be respected, according to traditional archivistic science rules and Italian archival laws.

For these reasons a system able to preserve already formed, arranged and structured archives, the noncurrent and historical ones, was conceived. They consist of records, no longer useful for their producer. DAX does not deal with current records and archives whose creation and management are delegated to the specialized systems. The current archives will be sent to DAX as they become non-current. Therefore, the task of this platform should be to maintain archives, in compliance with national and international standards, and in compliance with national and European archival law.

4. Technological Foundations

The Tuscan platform for the long-term preservation, as we mentioned, was devised using pre-existing organizational and technological infrastructures. Among those the telematic network, called RTRT (Rete Telematica Regionale Toscana - Tuscan Regional Telematics Network), plays a key role. It is a network with large capacity, spread throughout the region, connected to the Internet, and compliant to the national standards. The other fundamental infrastructure is the technological infrastructure for interoperability called CART (Cooperazione Applicativa Regionale Toscana - Tuscan Regional Applicative Cooperation). These enabling infrastructures comply with the Italian national legislation in terms of Public Administration standards. They have been certified and accredited at the national level, provide higher quality services than the standard market ones, realizing a multi-supplier model. Particularly for the entire Community they provide and ensure:

1. A set of connectivity services shared by the Tuscan public administrations;
2. Interaction with all the other subjects of Italian government connected to the Internet, as well as the networks of other institutions; they promote the delivery of quality services for citizens and private companies;
3. Shared exchange infrastructure that enables interoperability of information systems with external agencies;
4. The development of interoperable systems, according to the model of applicative cooperation, safeguarding data security, confidentiality of information, respecting the autonomy of the information assets of each administration and the current rules about privacy.

A particular relevance, in the context of the just mentioned infrastructures, is the implementation of interoperability¹⁵ among the different network actors. In Tuscany, we said, it is performed through a framework called CART. CART achieves the interoperability of applications of different organizations that decide to work together to get common supplying of public services. CART uses a set of software tools and defines a set of shared elements: management services, vocabulary, concepts, principles, policies, guidelines, recommendations and practices. These common elements become part of documents

¹⁵ “Interoperability within the context of European Public Services delivery, is the ability of disparate and diverse organizations to interact towards mutually beneficial and agreed common goals, involving the sharing of information and knowledge between the organizations, through the business process they support, by means of exchange of data between the respective ICT systems:” http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf.

arranging standards of interoperability among systems, and these documents are named “RFC e.Toscana”. Those standards are proposed by the technicians of the various domains and are open to public discussion. The negotiating process to produce interoperability standards involves public government, universities, research centers and private companies. These entities define the set of rules and specifications to ensure the interoperability of systems for a specific application domain. The standards take into account previous national and regional decisions or choices, as well as national and regional experiences. In particular, the Community recommends interoperability agreements, collaborates on the definition of services interfaces and on the process of accreditation of software systems and standards-compliant solutions (e.Toscana Compliance). The e.Toscana Compliance Committee, consisting of universities, research centers in Tuscany and local authorities representatives, ensures the governance of the process, by supporting the dissemination of approved standards, accrediting conformity of the software products with the standards and provides support to entities of the territory.

All this corresponds to the recommendations on EIF (European Interoperability Framework for the Pan-European e-Government Services)¹⁶ contained in the attachment “Towards interoperability for European public services” by the Social Committee and the Committee of Regions of European Commission. This last one suggests the creation and deployment of infrastructure to support interoperability. It also emphasizes the role of open standards and interfaces for the implementation of interoperability systems between applications and business processes related to e-government public services.

The sharing of interoperability standards for services means that the different administrations of the region operate in the same way approaching entities outside territory. On one side, this approach allows Regione Toscana to achieve full interoperability and governance issues; on the other side provides value for citizens and agencies, which perceive the Tuscan government as a single entity, rendering their access to services easier. These methods contributed to the growth of an eco-system of public services and created a culture of interoperability. The services can be submitted and proposed by any administration, and they are experienced as an opportunity by other members in the Community, to improve their services. The entire Community participates in the creation of a real eco-system; it benefits of coming out of new services, or, in case, of improvements or integration of services already existing. The emergence of new services is assisted by a process aiming at the full sharing of interoperability specifications.

5. Architectural Components

DAX is deployed in the regional territory through information architecture fully distributed coherently with the structuring of the regional network infrastructures. The central components of the platform are deployed at the regional data center called TIX (Tuscany Internet eXchange)¹⁷, which provides services to all administrations which are part of the regional Community Network.

Among the services that TIX data center provides, DAX uses:

- *Storage*, or rather the ability to extend incrementally the size of data storage and their backups. The chance of expanding the system is a key feature because of the impossibility to determine, a

¹⁶ See footnote number 3.

¹⁷ You can find news about all infrastructures on the site <http://www.e.toscana.it> where there are also links to sites specifically devoted to individual technical infrastructure.

priori, the maximum capacity of storage required. Another significant quality of DAX is its ability to store separately the archives of several administrations of Tuscany. This ensures that the different administrations retain their responsibilities on archives and their descriptions;

- *Disaster recovery*, which is the guarantee of saving information on different sites and the subsequent recovery from an unforeseen incident at a Data Center. This quality is essential in a system such as DAX aiming to preserve documents and information;
- *Operational or business continuity*, namely to ensure that DAX is able to operate even in the case of adverse events. This quality is desirable for a conservation system: access to information becomes strategic with serious emergencies.

The central components of the storage system at the TIX, receive packages of documents from the administrations of the territory. These send groups of documents from their applications by using the CART infrastructure that gathers, validate and submit them. The process of collecting, validating and transmission of packages takes place through standard device components of DAX. Those components, called Proxy-DAX and deployed on CART infrastructure, interact directly with the local applications of entities. The choice to implement specific Proxy guarantees:

- The distribution of the platform workload;
- The selected forwarding packages that are effectively to keep stored;
- Minimization of the use of network bandwidth between local applications and central components of the DAX;
- An effective security policy for communication between applications and DAX;
- The possibility of storing data and information in the proximity of each entity; this guarantees excellent response times in document and information retrieval;
- A significant improvement in quality of fault-tolerance of the system.

Another important task of the Proxy-DAX is to break large packages into smaller ones that will be reassembled by the central components of DAX. This option allows local applications to send virtually unlimited size packages.

The application interfaces made available, to the local software applications, by the Proxy-DAX are defined in appropriate RFCs e.Toscana.¹⁸ In line with the process e.Toscana Compliance, the RFCs e.Toscana concerning DAX were discussed inside the technical community, with the participation of many local and national companies, and finally they were approved by the e.Toscana Compliance Committee and have become regional standards. The four RFCs e.Toscana standard, about DAX, are technical documents that companies should consult to implement software and adapt their applications to the use of DAX.

¹⁸ The RFC e.Toscana , describing the application interfaces realized by Proxy DAX, are: n° 188 (<http://web.rete.toscana.it/eCompliance/portale/mostraRFC?idRev=682&idRfc=188>); n° 206 (<http://web.rete.toscana.it/eCompliance/portale/mostraRFC?idRev=681&idRfc=206>); n° 176 (<http://web.rete.toscana.it/eCompliance/portale/mostraRFC?idRev=642&idRfc=176>); n° 189 (<http://web.rete.toscana.it/eCompliance/portale/mostraRFC?idRev=629&idRfc=189>).

In addition to the access control decision by the document management service providers, DAX supplies and facilitates consultation to records and content information by the side of users in response to a request. The access has to take place using the application safely through an identification Smart Card.¹⁹ These cards are distributed to all citizens of the region: they are associated with a user profile and about that profile DAX combines its application roles. In practice this means that each person may have access to some features rather than others or have visibility of a portion of the archive and not others.

6. Archival Fundamentals and Standards

From the archival perspective DAX applies the principles of archival science and the rules of traditional archives arrangement and description, transferring them to digital archives. The basic principle from which the analysis took the place is that the archive of an administration is an *unicum* and all documents, produced in the history of that administration, are part of this unitary system. The platform describes and manages digital documentation but it describes also traditional paper records (or more generally analogue objects). We thought this was a good way to ensure the uniqueness of the archive according to its provenance. At the same time, we kept in mind that preservation and description of digital records and archives require to highlight the peculiarities and keen differences between the two worlds.

In the first place DAX provides a solution to two problems that, although closely related, are not completely comparable:

1. The long-term preservation of digital records;
2. Archive management—depository and historical level—both analogue and digital.

The choice to manage, through this system, the hybrid archive, and not only digital, follows from the fact that the archives and single practices of our administrations are still largely produced on paper.

The platform DAX is based on the ISO OAIS (Open Archival Information System),²⁰ and in compliance with OAIS standard focuses on the preservation of information packages. In addition it describes the documental and archival context of creation, conservation and preservation, and identifies an application area. The decision to build a system OAIS compliant depends on we are in agreement with principles and topics of the standard, and we have considered the standard next to our reality: a community of well-defined baseline, share knowledge, standardization well tested tools, languages and methods. In fact it provides theoretical and interpretive trends not only for archives and its forming objects, but also for their context, and it suggests organizational answers.

With regard to relations with the environment, three types of entities or systems interface DAX: “producer” that creates and sends to the system the records and archives to be preserved, “user” who consults the archives by distance, in space and time; “manager” who, structured into several kinds with different responsibilities, takes charge of the management, maintenance and updating of the system as a whole. And finally, the objects to be preserved.

In this initial phase of the system’s use we decided to begin maintaining two broad categories of objects: general administrative records and health records. According to the several stages that OAIS

¹⁹ It is a hardware device, similar to a credit card, which contains all the information related to the digital certificates of the subject and which allow a certain authentication.

²⁰ Open Archival Information System is the name of the standard ISO: 14721:2003 which defines concepts, models and functions related to digital and aspects of digital preservation.

contemplates, records are processed and aggregated in packages, logical containers of records and information about records and their preservation. The packages distinguish themselves depending on the stage and type of entities (respectively called SIP and AIP)²¹ (Figure 1).

The system stores records in several digital formats that are made known to the community through a “list of allowed electronic formats” that can be expanded over time as needed. DAX also intends to keep any type of documents and files: text, database, image, e-mail or e-mail archive, map etc.²²

DAX has all the features of feeding, managing and finding of intermediate and historical archives, as regards both paper and digital objects, and it manages:

- *Ingest process* from the creation to preservation phase by receiving packages of documents and their metadata, from document production systems. The metadata sets forming the packages for the ingest process and keeping, are compliant with the national Italian standard UNI-SInCRO.²³ The ingest process occurs through the acquisition of “packages” (SIP-OAIS) consisting of a number of archival aggregates²⁴ and / or single record.²⁵ As DAX manages the intermediate and historical archives, these packages must contain no-active archival units and closed files or, at most, single non current record;
- *Logical organization and description* of the archives, in compliance with international standards for archival description.²⁶ The policy of the system is that the records description, required at the moment of ingest, is a dynamic description, enriching during the phase of preservation and at the request of access. The metadata sets, to describe and holding archives, were enhanced and compared with the sets worked out numerous international research projects, we held out as a model.²⁷ One of the great efforts of this project was, indeed, to try interpreting standards and best practices, with the aim to exploit the rich outcomes of many international projects;

²¹ We are speaking about the packages provided by the OAIS. There are three different types of information package: SIP (Submission Information Package) that is made at the time of ingest to the archive by the creator, the AIP (Archival Information Package), for storage in the DAX, the DIP (Dissemination Information Package), which is composed with the data relating to the distribution and access.

²² It was decided, in the phase of the activation of DAX, to start the system on certain types of archives and then specific formats, but through small changes and stepwise refinement, based on the needs of the government, will expand the range of sizes and the types of storable.

²³ This is the Italian national standard UNI 11386: UNI 11386:2010 - interoperability support in the Storage and Retrieval of Digital Objects (sync). It is the result of the National Italian unification (UNI), which was established within Subcommittee DIAM/SC11 (Management of archival documents), in 2009; it was a special working group, called Synchron.

²⁴ This archival unit is a consistent set of documents, grouped by a person for the purposes of its business, according to the common reference to the same subject.

²⁵ See footnote number 9.

²⁶ We are speaking about ISAD (G) (General International Standard Archival Description) and ISAAR (CPF) (International Standard Archival Authority Record For Corporate Bodies, Persons and Families) are standard adopted by the International Council on Archives, in order to define unique tools for the description of archives, for the registration of documents produced by organizations, individuals and families (www.icacds.org.uk/eng/). The first edition was published in 1994. These set of metadata are returned to the user organized according to the model EAD (Encoded Archival Description) and EAC (Encoded Archival Context of). Developed and published (1998) by the Society of American Archivists in partnership with the Library of Congress for encoding tools appropriate archival.

²⁷ Among these especially InterPARES - The International Research on Permanent Authentic Records in Electronic Systems (<http://www.interpares.org/>), PREMIS - Preservation Metadata Implementation Strategies

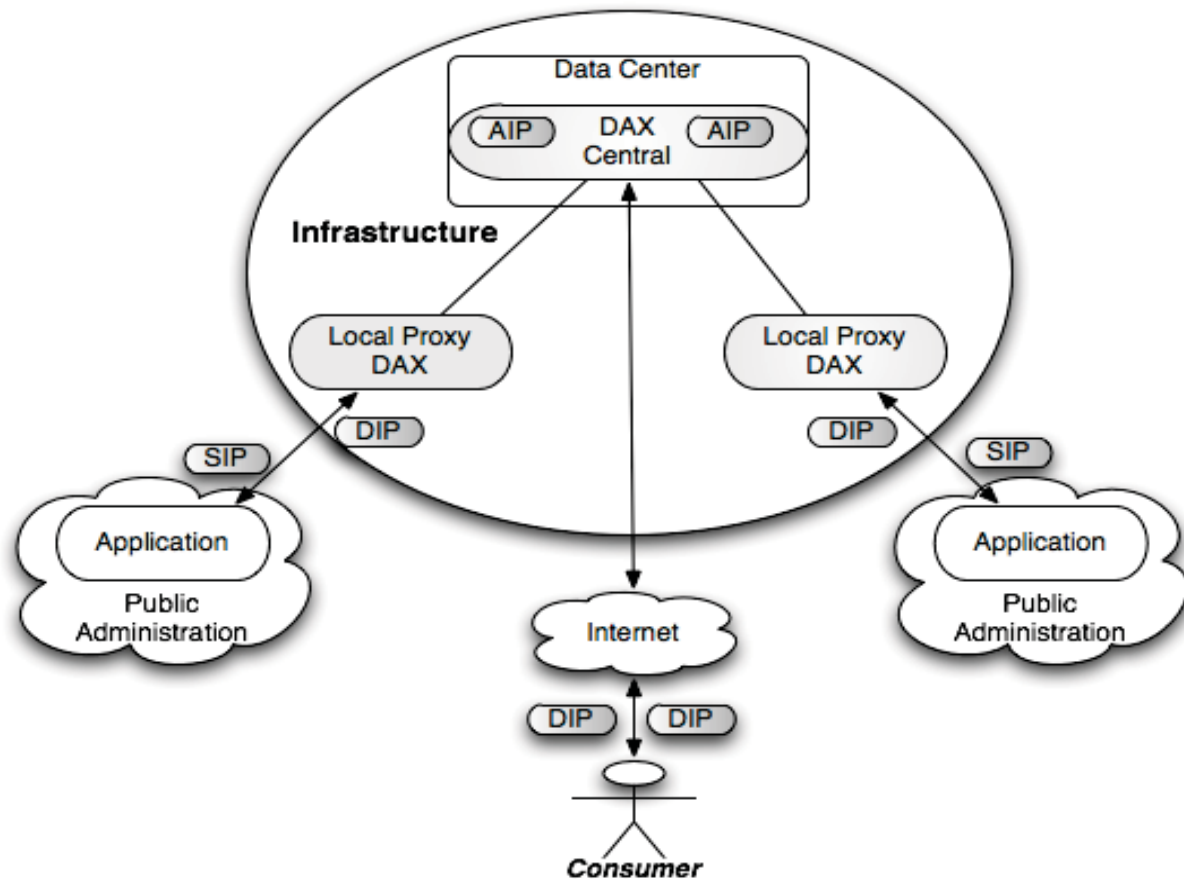


Figure 1. DAX architecture overview.

- *Appraisal and selection* and subsequent retention of predestined records; this phase includes implicit request for authorization by the Ministry of Cultural Heritage, in accordance with Italian law;
- *Access and rendering* of records and aggregated records and files by internal and external users of the system;
- *Access and consultation* by the Auditor (e.g., State offices: Ministry of Cultural Heritage – Soprintendenza Archivistica, Ministry of Innovation etc.) or justice organization;
- Clear definition of roles and responsibilities.

From the point of view of technological capabilities, the platform preserves digital documentation and, in order to cope with obsolescence of technology and software, enforces a continuous activity of control and migration.

(<http://www.oclc.org/research/activities/premis-rlg.html>), METS - Metadata Encoding and Tradition Standard (<http://www.loc.gov/standards/mets/>).

Furthermore DAX has another very important feature: it produces a complex auditing system that keeps the memory of all the logs occurred, both automatic and manual. This is an excellent method to control the functioning, processes and reliability of system, to preserve memory of operations and to assure safety of the archive. Based on this auditing system, as well as a continuous check on the data, it will be possible to make an assessment of the procedures put in place for storage. The audit data, obtained and organized by all the features of the platform, allow to measure the compliance of processes and procedures with respect to the characteristics of the system and their application, with reference to what has been defined in the analysis as a guarantee of reliability of the system and what is required by the certification systems.

7. The Architectural Features of DAX

An initial choice has set that the system was articulated into two connected but independent subsystems, playing different but complementary roles, since the system is logically unique. On the one hand, we have the part which governs the real archive, that focuses on the organizational and use aspects, carried out in accordance with the OAIS reference model. On the other hand, a complex storage that focuses on aspects of the Italian law topic called “conservazione sostitutiva”.²⁸ This kind of preservation is a legal and technological procedure that is regulated by the Italian law, to ensure, over time, the legal validity of an electronic document. According to the current laws, the digital document is “locked / closed” in form and content through the digital signature and time stamp, which, by setting the exact date and time of its crystallization, anchor it temporally and guarantee the fixity of information.

Thanks to this division into two subsystems, platform DAX is able to satisfy two requirements that initially seemed difficult to conjugate:

- Preserve digital objects “freezing” by hashing techniques and asymmetric key cryptography, according to the Italian law (bit-preservation);
- Preserve packages formed with records and metadata, assuring they can be changed over time (package-preservation), according to the OAIS reference model.

The system, as a whole, is responsible for ensuring that, in medium and long-term, records retain: integrity and authenticity, accessibility—as long as needed—and availability, legibility and intelligibility, and reproducibility.

Each of the two subsystems has some peculiarities but in general, given those guarantees, DAX takes charge of:

- Implement, manage, historicize tools and data about the organizational, archival, procedural and technological context. These contextualization is functional to the description of the stored documentary heritage: organizational structure of producer, classification plans and indexes, content types, appraisal plans, vocabularies to interpret specific metadata, encodings or terms, associated with documents;

²⁸ It is a set of rules laid down by a decision of the Authority for Informatics in Public Administration (AIPA) of 2004, n. 11. This rule is in the process of substantial change and the changes can be found at <http://www.digitpa.gov.it/gestione-documentale>.

- Provide services and references to search and browse the preserved documentation;
- Prevent the obsolescence of hardware and software through continuous adjustments and through processes of migration of digital stored documents;²⁹
- Record and store—in a unique audit system that returns summary data queried at multiple levels of detail—the tracking of each access, change and activity on the system (access, technological changes, and updates the metadata of digital documents);
- Keep alive the digital signature, in compliance with the Italian law (by the application and renewal of timestamps or by logging ingested package into the system);
- Receive, update and maintain metadata about the document in the archive.

As mentioned several times, the platform is designed as a system of long-term preservation at the service of the Region of Tuscany, but also local government of the community. Which is why both of the two subsystems are designed as multi-entity, i.e., the only installation maintains complete logical distinction, even if it can manage archives and contextual archival metadata of several entities.

8. Governance

Such a many-sided system requires a governance able to ensure effective control and management. The system is complex in nature because of the role that it aims to perform, because of the plurality of treated subjects, the many different involved responsibilities, the complexity of functions at stake.

Consequently, the working group on DAX focused on what would be the best form of government of the system, moving from the organizational context of departure. A strong point was that the system had to be co-managed by the entities that produce the archives with a direct involvement in the Community. Every administration would have to retain full responsibility for its archive, even if the system has to be managed and conducted through a board of expert with appropriate skill. This board has to be a qualified and a recognized group, equipped to manage digital preservation system, supervising it from every point of view: organizational, technological, legal, political and about diffusion process. A group able to provide appropriate security guarantees, effectiveness of technology, accessing to adequate technical equipment and to professional training.

Since the beginning of the project, a number of activities and meetings have been scheduled, to support different subjects involved in the management of platform DAX. The idea was to encourage public administrations and their technicians to discuss and test the system according to their need, experiences, and resources.

These activities are coordinated by a highly specialized team, that has entrusted the management of the system, called Centro di Responsabilità per la Conservazione Digitale (CRCD) (Responsibility center for digital preservation). The CRCD ensures the process of long-term preservation and, especially, it takes charge of the definition of meta-information and archival rules. The board has the responsibility for the necessary adjustments of the system to international standards of long-term preservation, and for the

²⁹ In particular, CNIPA - Centro Nazionale per l'Informatica nella Pubblica Amministrazione, Deliberazione 2004, 19 February, n. 11, "Regole tecniche per la riproduzione e conservazione di documenti su supporto ottico idoneo a garantire la conformità dei documenti agli originali," art. 3 and 4.

updating of the system to any possible change in the Italian legislation. The CRCD is also in charge of the accreditation of DAX to the National Agency for the Digitization of Public Administration, as well as of the possible subsequent adjustments of the system to their requirements or to new international standards. The CRCD ensures consistency and compliance with safety plans, also in respect to privacy policies. Finally, as regards new applying institutions, the CRCD coordinates the deployment process and the activation of organizational process, as well as its operative start.

Four are the institutions which play an important role in the management of the DAX platform, each one with clear responsibilities: Soprintendenza Archivistica per la Toscana, Regione Toscana and the ICT companies involved in the project. In this respect, the CRCD encourages and coordinates the relationships among all these entities so that they can operate together consistently, sharing a common goal. Among these four, Soprintendenza Archivistica per la Toscana plays a crucial role, as it is an office of the central State which, according to the Italian law, is in charge of monitoring and protecting historical archives of all public bodies, disregarding their political level. Further relevant institutions participating in the project are, as mentioned, Regione Toscana and the ICT companies which deal with the functioning of the technological components, according to the directions issued by the CRCD. A fifth fundamental partner adds to these four basically stable partners: the public institution that decides to use the DAX platform for the long-term preservation of its archives.

In order to have all these entities operating consistently under a defined articulation of tasks, a formal subscription of an agreement is required. This agreement states roles and responsibilities of each partner, as it also states the rules for the common governance of the system and the compliance guidelines for interoperability. Further relevant documentation is the technical documentation provided for the activation and management of DAX, the manual for conservation, standards and the necessary repertoires: all of this proves to be crucial for digital preservation of archives.

The Road to Providing Access to Kenya's Information Heritage

Digitization project in the Kenya National Archives and Documentation Service (KNADS)

Francis G. Mwangi

Kenya National Archives

Abstract

This paper discusses how the Kenya National Archives and Documentation Service (KNADS) started digitizing its records in its quest to give access to millions of documents in its holdings. KNADS under the law has the responsibility of taking "all practicable steps for the proper housing, control and preservation of all public archives and public records in Kenya." (Cap 19 Laws of Kenya) The paper indicates that the purpose of the programme is twofold, to give access to the information contained therein and to preserve the original archival materials for posterity. The paper shows the work plan that KNADS adopted in digitising a part of its collection, it shows the different methodologies adopted by KNADS to achieve these objectives, and the challenges it has faced so far. It concludes by indicating that although considerable progress has been achieved by KNADS in ensuring that the most consulted records in its collection have been digitalized, there is a need to look for new methods of achieving its goal in a shorter period as well as giving access to those records it has digitized.

Author

Francis G. Mwangi was born in Kenya in 1967; currently he is the acting Deputy Director of the Kenya National Archives and Documentation Service, Administration and Finance. He is also the head of the Films and Audiovisual Records in the Kenya National Archives. He is a holder of a Master of Science Degree in Archives and Records Management from Kenyatta University, Nairobi.

1. Introduction

KNADS is a Department in the Ministry of State for National Heritage and Culture in Kenya. It was established by an Act of Parliament in 1965 to take all practicable steps for the proper housing, control and preservation of all public archives and public records of enduring value, and make them available for public access. (Public Archives and Documentation Service Act, Cap 19 Laws of Kenya)

The weight of this responsibility is rapidly changing in the current digital environment and has become a real challenge. It's worth noting that some of the archival records and media are old, brittle, and delicate that requires careful handling. It's therefore important that the Kenya National Archives and Documentation Service actively intervene to ensure that these records will be available today and in the future. The use of archives is the goal that all archivists would endeavor, but the availability of archives for use by the public, and indeed all other aspects of archives management depend on archives being properly preserved and cared for, and now being made available to the users all over the world through the techniques of today and therefore the need to digitize our records.

Traditionally archivists have shaped their preservation activities around the notion of Permanence; their objective has been to ensure the permanent preservation of archives. This is despite the fact that no record, no matter how well protected and cared for, enjoys an unlimited lifespan. Internal processes of decay ultimately defy even the most sophisticated intervention by archivists.

Since 2007, KNADS has been carried out a digitization programme. This has involved digitizing some of its oldest and heavily used archival materials, some dating back to more than 100 years. This

programme has resulted in the digitization of close to 12,000,000 documents of archival records (KNA/8/2 Vol.11). While this may sound a big number, it is a very small portion (3%) of the more than 400,000,000 pages of archival materials in the custody of KNADS. It should also be noted that digitization is both capital and labour intensive as will be shown by the paper.

2. Background

In 2007, the Kenya National Archives and Documentation Service (KNADS) decided to start digitizing the records of Coast province that are held at its Nairobi headquarters. They were estimated to number slightly over 1.7 million pages of both bound and loose pages. These records were selected because of their age and the heavy usage by clients; it was argued that reformatting them into digital copies would allow wider use and ease of access while preserving the original. Constant handling of these records by users over the years has rendered them to wear and tear. Continued usage of the original document would lead to their destruction. By digitizing these records, the original would safely be preserved while the digital surrogates would be used for access. By virtue of them being digital, it would be possible to produce surrogate, and derivative files without any damage to the original digital master. The digital object would then retain all significant information contained in the original document(s), and under appropriately stringent conditions related to migration, refreshing, and backing –up of the original file, it should survive over time.

2.1 Ease of Access

Demand for access to original materials, often termed as ‘primary source materials’ is increasing every day. Members of the public often require instant access to records. Archival records are mostly single copies and as such they cannot be accessed by multiple users at the same time. However, with digitization, it will be possible to avail these records to the users online where multiple users can access them. The trend worldwide in archival institutions has been to digitize and even make these records available in the internet where users can access them after paying a stipulated fee.

It is envisaged that the digitization will be undertaken at a very high archival-master level quality that will allow for multiple output (e.g., print, microfilm, access images, thumbnails, etc.) when need arises.

2.2 Baseline Assessment

A baseline assessment was carried out by the department to establish the nature of the source materials. About 30% of the records were found to be very fragile and brittle and therefore, they required special handling and equipment when scanning in order to avoid any damage. About 10% of the documents required wide format scanners as they were not of standard size. Big portions of the document were hand-written and therefore, Optical Character Recognition (OCR) would not be performed on them.

2.3 Project Outcomes

After these materials have fully been digitized, users will no longer need to access the original materials unless on very rare occasions to satisfy curiosity; for authenticity and legal purpose. It will also be possible to do full text searches on the records as they will be indexed and as such it will be easier to

automatically identify documents containing relevant information, something that is not possible with paper records.

3. Progress of Implementation

The first three years of the digitization project was outsourced to private firms. This was out of the fact that the department did not have internal capacity in terms of adequate skilled manpower and equipment. This period saw the digitization of over 10 million pages of archival records. The contracted firm used to have a workforce of over 40 people working in two shifts (day and night). In the last two financial years, the department decided to undertake the project in-house with view of developing internal capacity for sustainability. The department procured twenty computers, twenty medium duty scanners and one wide format scanner. The intention was to use the department's staff to undertake the exercise.

4. Roadmap to Digitization of the Records of Coast Province in the KNADS

4.1 Work Plan and Payments

Milestone one: Digitization of Coast Province Documents			
OBJECTIVE	RESPONSIBILITY	MILESTONE/INDICATOR	BUDGET (KES)
Pre-Digitization Accession 11 - 22 June 2007	Outsourced Company	1,685,000 pages unclipped, dusted, demagnetized and batched at KES. 0.50	842,500.00
Scanning 22 June – 10 August 2007	Outsourced Company	1,685,000 pages digitally captured, OCR. edited and indexed, KES. 2.00 per document.	3,370,000.00
Post-Digitization Accession 22 June – 15 August 2007	Outsourced Company	1,685,000 re-clipped, transported to storage and catalogued at KES. 0.25	421,250.00 100% payment on completion

4.2 Workflow

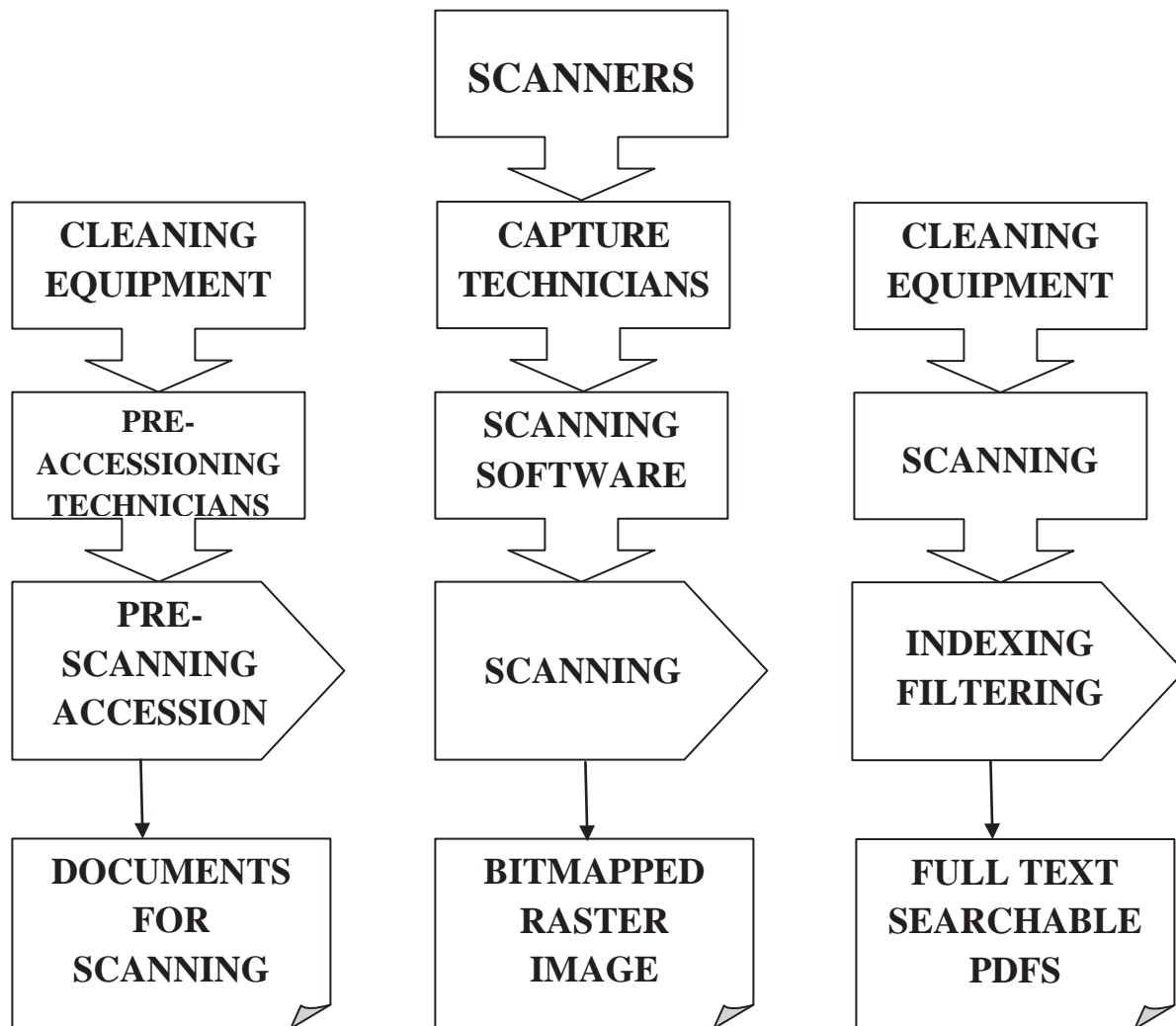


Figure 1. Workflow and events leading towards searchable document in KNADS.

5. Methodology

The first step was the preparation of the 1,685,000 documents through cleaning. Demagnetizing, unclipping and batching that mirror the output indexing and metadata.

The second step was scanning with various equipments according to documents specifications. All the originals 1,685,000 documents were scanned into digital master documents in TIFF CCTT4 format (resolution 6000 pixel – 8 bit greyscale) with lossless compression. The raster files were then saved into backup media repository.

The third step involved filtering the documents into the required formats. Filtered copies from the digital masters were made according to the needs. PDFs were done with metadata and full-text search

capabilities for retrieval done through OCR and documents schemas while GIF and JPEG thumbnails were created for intra/internet display. The work was divided into 5 objectives as shown below

5.1 Objective One: Pre-Digitization Accession

Input: 8 members of staff, vacuum cleaners and demagnetising hardware

Activities: 1. De-magnetising the documents
2. Dusting
3. Unclipping documents
4. Batching

Output: 1,685,000 clean unclipped and batched documentsin

Timeframe: Two weeks

5.2 Objective Two: Scanning

Input: 4 capture/editing technicians, scanners, computers and softwares

Activities: 1. Scanning
2. Editing
3. Indexing

Output: 1,685,000 indexed raster images

Timeframe: Eight weeks

5.3 Objective Three: Indexing and filtering

Input: 6 capturing/editing technicians and sorting staff

Activities: 1. Optical Character Recognition
2. PDF Conversion
3. Document Schemas and Metadata tagging

Output: 1,685,000 text/metadata searchable PDFs

Timeframe: Four weeks

5.4 Objective Four: Post Digitization Accession

Input: 6 re-clipping casuals

Activities: 1. Re-clipping documents
2. Transporting Documents
3. Arranging/Cataloguing documents

Output: 1,685,000 catalogued and achieved hard copies

Timeframe: Two weeks

5.5 Objective Five: Installation of Hardware

Input: 2 hardware & software engineers

Activities: 1. Installation of backup external disks
2. Installation of Server

Output: Configured Storage Hardware System

Timeframe: One week

5.6 Scanning Hardware Used

- 2 high duplex scanners capable of scanning 80 pages per minute, but given that these are archival document, the output is much less than that.
- 4 book scanners one being planetary capture scanner
- 1 wide format scanner (A0 size)

Storage Requirement: Master Documents – TIFF: 1685,000 pages x 2.4 megapixel = 3,840,000 megabytes = 3,840 gigabytes = 3.84 terabyte of space required

Filtered Compressed PDFs: 1,600,000 pages x 120kb

Software requirement: to manage the repository for retrieval, Adobe Acrobat and Acrobat Reader are being used.

5.7 Quality Assurance

A team of officers from KNADS were detailed to be carrying out quality control analysis at various stages of the project to verify that all reproduction is up to standard. The quality assurance analysis was carried out on random sample of 10% of all the stages. The quality assurance analysis were to determine the following:

- Size of image
- Resolution of image
- File format
- Image mode (i.e., colour images are in colour, not grayscale)
- Bit depth
- Details in highlights and in shadows
- Tonal values
- Brightness
- Contrast (e.g., stark black and white contrast on anything except simple line drawings)
- Sharpness
- Interference
- Orientation
- Noise
- Alignment of colour channels
- Cropped and border areas, missing texts, page numbers, etc.
- Alignment of images
- Missing lines or pixels
- Text legibility and meta data capture

6. Challenges

The following are some of the challenges faced during or immediately after finishing the scanning exercise.

- There were 1,585 wide format drawings and maps which had not been captured initially and which form part of the records of coast province. Since the drawings and maps are an integral part of the contents of the repository, these drawings and maps have to be scanned and according to the firm we had outsourced, the cost will be at KES. 150.00 per drawing/map, this was because the maps and drawings would require wide format scanners, after shopping around we found that the prize was actually KES 50.00 less than the market price.
- We also realized that there was another consignment of 190 boxes in the basement repository with 437,000 extra records of the same provenance, i.e., Coast Province which had not been captured in the initial count as they were in another repository. The cost of digitizing them was charged as the earlier records.
- We also needed to store, make backups and manage the electronic repository created with two 500GB external hard disks, one hundred and twenty 4 GB DVDs and PDF archival software, all totaling to an extra KES. 581,500.00
- Quality: while the contractor has been given the specs for digitizing the records, where mass digitization is taking place, quality may be compromised, especially given the fact that archival quality digitization requires a high resolution. This meant that an archives staff has to be within the process all the time. The work has to be checked again and again to ensure the right image is achieved always.
- Expenses: As will be seen, there are always “other” things and issues that crop up during or after the completion of a project. Therefore, apart from the original budget, extra funds have to be kept aside for such issues and things that are bound to arise.
- Consistency of the filing system: it was noted that during quality control, it should be ensured that the order in filling of the manual documents is maintained in the digital document, the documents in every file should be scanned from back to front just as the file has grown., where a document has more than one folio, the scanning should start from page one of the last page and in order to ensure that this was done properly, the departmental committee dealing with monitoring and evaluation of digitization programme was to closely and regularly visit the site and assess the progress.
- Re-boxing and re-shelving of files: this is a major challenge during a digitization exercise especially where the services have been outsourced. The archives developed a stamp having the inscription ‘verified’, ‘date’ and ‘signature’ and the re-shelving team had to look and ensure that the records have been digitized before they are returned to their boxes.
- Outsourcing: While it is appreciated that outsourcing resulted in the digitization of over 10 million documents within three years, the exercise was very expensive and yet there was no specific budgetary allocation for the project. This meant that the project could not be sustained through outsourcing.

- **In-house Project:** The initial intention was to deploy officers from other sections to carry out the digitization on a rotational basis. The officers could not afford to dedicate adequate time for the exercise since they already had other duties to take care of in their respective sections. The department, therefore, resorted to hiring of casuals to assist staff in the digitization exercise. The money allocated during the last two financial years for hiring of casuals is a cumulative total of 3 million shillings. This figure has only been adequate to hire 20 casuals for a total 9 months. This implies that half of the year no digitization takes place yet the total number of people undertaking the exercise is only half what the external contractors used to have. Furthermore, considering that the programme is long-term, it is also not sustainable to use casuals year-in year-out.
- **Access:** although the primary reason for digitization was to offer access and preservation of the original, the actualization has been slow. This is because of the following reasons:
 - **Website:** The Kenya National Archives and Documentation Service website had been hacked and it took time constructing another one.
 - **Content:** hosting the website with all these data content became an issue, the capacity of most of our service providers was low and as such they could not handle our request.
 - **Technological Challenges:** there was no one who seemed to know exactly how to go about uploading the material that we wanted uploaded without compromising the rest of the material in the server.
 - **Payment and Charges:** if we were to charge for the material accessed, how would we go about it and what modes of payment would be accepted, noting that this is a government department.

7. Way Forward

Long-term solutions to the digitization project needs to be found with a view of ensuring that it is sustainable. These would include the following:

7.1 Personnel

To clear the huge backlog of 97% of un-digitized materials, the department would require engaging 200 staff on a full time basis. That number of staff would be able to digitize 200,000 documents per day at the rate of 1000 per person. At that rate, the department would be able to clear the backlog in five years.

7.2 Training

Digitization is a technical area that requires professional skills to ensure quality output. It is therefore important for all the staff engaged in the project to undergo thorough training in digitization both at middle and advanced levels.

7.3 Equipment

While the department has already acquired digitization equipment, these would not be adequate if the department was to acquire the desired number of staff. It would therefore be necessary to acquire more

equipment. Additionally, ICT equipment becomes obsolete very fast. It is therefore important to budget for the upgrading and replacement of the equipment that are no longer serviceable.

7.4 Budget

As noted earlier, digitization is a capital intensive project and as such the department needs to introduce a specific budgetary item for this purpose with adequate allocation to cater for its needs.

7.5 Partnership

Another way of going about the digitization programme is to partner with stakeholders and friends of archives wherever they may be, the partnership can be in form of assistance, technical knowhow, equipment, financial or any other way that can be offered to make the programme a success story.

8. Conclusion

Since archival material exist in single copies, they had to be handled with utmost care. For this reason automatic document feeders were avoided, unless the papers were in a very good physical conditions and of standard size. Another rider adopted was re-assembling back of all the documents into their respective files and the cover of the file stamped 'digitized'. Finally, since the files are read from back to front, digitization had to be done in the same way with folio one starting off till the last folio to be filed becomes the last.

Though it had been hoped that by undertaking the exercise in-house the challenge of sustainability would be resolved, this has not been forthcoming as the department is too thin on the ground in terms of personnel. Hiring of casuals was an appropriate stop-gap measure but it is also not sustainable in the long run. It is, however, worth noting that the department now has adequate equipment for digitizing records. However, the challenges of internal capacity need to be resolved with long-term solutions that would make the project sustainable. The international partners and friends are invited to assist the archive move a step further than where we are by making what we have already digitized available to the world.

A Preservation Infrastructure Built to Last

Preservation, Community, and HathiTrust

Jeremy York

Project Librarian, HathiTrust

Abstract

This paper describes the strategies HathiTrust is taking to build a collaborative infrastructure capable of ensuring long-term access to digital collections at scale. HathiTrust's approach recognizes the deep interplay of social and technical factors that support our collections, and will determine their persistence and availability over time.

Author

Jeremy York has been the project librarian for HathiTrust since July 2008. His primary duties include project coordination among the partnership, maintenance of HathiTrust's informational web site, and activities surrounding new partners and partnership contracts. Jeremy received a bachelor's degree in history from Emory University in 2001 and a Master of Information Science from the University of Michigan in 2008, with a specialization in archives and records management.

1. Introduction

HathiTrust is a partnership of academic and research institutions that are pooling resources to collaboratively preserve and provide access to the cultural record. The core of the preservation strategy centers on a digital repository that is owned and operated by the partners to ensure the long-term preservation of digital materials owned by their institutions, and facilitate access to the greatest degree allowed by law or third party agreements. The repository was launched in 2008 and currently contains more than 10 million volumes, making it one of the largest research library collections in the world. This paper offers insights into the guiding principles and ideas that underlie the repository, and specific strategies the partners are employing to preserve and provide access to digital collections at such a scale.

2. Setting

In the last 10 years, the time in which HathiTrust was conceived and initiated, there has been an explosion in the amount of materials digitized and produced digitally by libraries and other cultural heritage institutions. This has resulted in an increased focus in the cultural heritage sector on issues of digital preservation. Libraries and other institutions have grown significantly in their knowledge of the specific components involved in digital preservation, such as formats, media, and management of digital objects over time. They have grown also in their understanding of, and tools for evaluating, attributes and characteristics of "trustworthy" initiatives for long-term preservation. The challenges of preserving our digital present and past have been increasingly well defined. However, questions remain about the best ways to meet these challenges, from preservation models to employ (e.g., distributed versus centralized architecture), to formats and technologies to use, to specifications and best practices to follow.

In seeking to address these challenges, and in building a preservation infrastructure designed to operate at tremendous scale, HathiTrust has taken an approach that recognizes preservation first and

foremost as a social and collaborative activity. This approach has led to technological, architectural, and procedural decisions that, while important in their own right, are subordinate to, and guided by, an overall aim to meet the needs of a targeted community, even as the needs of that community change over time. This paper walks through the specific strategies that HathiTrust is taking to address common challenges in digital preservation, including issues of authenticity, reliability, scalability, sustainability, and discovery and access, in light of two guiding principles: that it is we, collectively, who are responsible for ensuring the persistence and availability of our cultural record; and that we can do more together than we can do separately.

3. Community

Viewed developmentally, the problem of preserving digital information for the future is not only, or even primarily, a problem of fine tuning a narrow set of technical variables. It is not a clearly defined problem like preserving the embrittled books that are self-destructing from the acid in the paper on which they were printed. Rather, it is a grander problem of organizing ourselves over time and as a society to maneuver effectively in a digital landscape. It is a problem of building—almost from scratch—the various systematic supports, or deep infrastructure, that will enable us to tame anxieties and move our cultural records naturally and confidently into the future.

(Task Force on Archiving of Digital Information 1996, 7)

The quote above is taken from the 1996 report of the Task Force on Archiving of Digital Information—a report foundational to the establishment of criteria for certifying trustworthy digital repositories, and significant in the development of the framework for Open Archival Information Systems.¹ One of the primary focuses of the report was on the need to advance the establishment of trusted systems for preserving digital information (1996, 9-10). As the quote above demonstrates, the report also recognized the social framework and interplay of social factors that both support and benefit from trustworthy digital preservation.

The importance of both of these aspects, the technical and the social, are carried forward in the OAIS model and in Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC), the culmination of many years work to establish criteria for certifying trustworthy digital repositories. It is significant, however, that social components in both of these models are articulated primarily in a service-consumer relationship. The OAIS model defines a repository's Designated Community as "an identified group of potential Consumers" of information, where Consumers are the "persons, or client systems, who interact with OAIS services to find preserved information of interest and to access that information in detail" (Consultative Committee for Space Data Systems 2002, 1-8). The Designated Community is specified separately from a Management function, which typically acts to provide funding, conduct reviews of the OAIS, determine pricing policies, and "provide support for the OAIS by establishing procedures that assure OAIS utilization within its sphere of influence" (2002, 2-8).

In TRAC, one of the three overarching areas comprising the evaluation is organizational infrastructure, including issues of governance, staffing, finances and sustainability, contracts and liability,

¹ The report was the source of the recommendation to institute a dialogue on the "standards, criteria and mechanisms needed to certify repositories of digital information as archives" (1996, iv), and is taken as a point of reference in the Trustworthy Repositories Audit and Certification: Criteria and Checklist (CRL and OCLC 2007, 1). It was also the basis for the Preservation Description Information in the OAIS model (Consultative Committee for Space Data Systems 2002, B-1).

and succession (CRL and OCLC 2007, 3). TRAC explicitly recognizes the social elements that underlie a trustworthy repository. TRAC borrows the definition of a Designated Community from OAIS, however, reasserting from an earlier 2002 report (RLG and OCLC 2002) that “the definition of a trusted digital repository must start with “a mission to provide reliable, long-term access to managed digital resources to its designated community, now and into the future” (CRL and OCLC 2007, 3). Here too, the envisioned relationship is weighted toward one where an organization (stakeholders fulfilling management functions) provides services to a separate body of end users.²

The notion of Designated Community does not exclude the idea or possibility that designated communities and stakeholder communities could be the same, but it is worth affirming this possibility explicitly, as it is precisely the model under which HathiTrust was established and operates. The partnership is a community of academic and research institutions that are collaborating to provide a shared digital preservation infrastructure that will enable them to better achieve their goals in provisioning the cultural record for the advancement of scholarship at their institutions. By doing this, the partnership serves immediate access needs of end users who are part of this community (those engaged in scholarship and research at the partnering institutions, including, as they may be, stakeholders, staff, students, and faculty, etc.). HathiTrust’s Designated Community, then, encompasses both those who steward and manage the digital archive, and the immediate users of information contained within the archive. As the partners are committed to using their collaborative services to produce public goods (making materials in the digital repository as open and available to anyone in the world as possible), they are able to reach beyond their designated community to serve a broader worldwide audience of libraries and library users.

There are clear strengths to this arrangement, where there is a tight coupling between those who support and manage the archive, and those who use and benefit from it. Perhaps the most important of these are first, that it provides the basis for a deep, collaborative social infrastructure where institutions are able to leverage common interest, distributed expertise, and diverse resources to achieve common and institution-specific goals more effectively and efficiently. Second, it creates strong forces that favor long-term sustainability. The sustainability of the archive depends on the ability of shared management and governance to ensure the archive continues to benefit the investing partners, and on the continued interest of the community in general in supporting scholarship and research. Both of these are fundamentally social factors.

A key element in ensuring the archive’s ability to benefit those supporting it, however, is the technology used and the archive’s technological approach in general. As the 1996 Task Force report noted, “We can afford to continue and increase economic and social investments in digital information objects and in the repositories for them on the information superhighway if, and only if, we also create the archival means for the knowledge the objects and repositories contain to endure and redound to the benefit of future generations” (1996, 9-10).

The remainder of this paper describes the approaches HathiTrust has taken to address challenges in preserving and providing access to digital information at scale in light of the social factors that ultimately underlie its success and sustainability. The partners have striven to develop robust technological infrastructure that is designed above all to be responsive to community needs for preservation and access, and that prioritizes meeting these needs over the long-term, even as technologies and implementations change over time.

² Examples given in OAIS of Consumer interactions include “questions to a help desk, requests for literature, catalog searches, orders and order status requests” (Consultative Committee for Space Data Systems, 2-9, 2-10).

4. Overarching Considerations: Scale, Preservation and Access, Openness

There are three broad considerations that have had a significant impact on the design and implementation of the HathiTrust repository. All result from underlying goals to meet the needs of the designated community. These considerations are the exceptional scale of the repository, a philosophical belief that the value of preservation is gained through access—that there is no value to a community of preservation without access, and a strong commitment to openness. Ultimately, HathiTrust’s need-based approach to the development of services and strategic directions is one of its strongest attributes.

4.1 Scale

In his testimony as part of the Google Settlement fairness hearing in 2010, Paul Courant provided a summary of the purpose of the libraries that were engaging in large-scale digitization of their collections in partnership with Google. The excerpt of his testimony below characterizes well the primary needs of HathiTrust’s designated community:

Without reliable access to the scholarly record, we cannot know what has been known, what has proved fruitful and fruitless in the past. The broad social benefit that derives from the progress of science and the useful arts depends on the ability to find, use, and reuse the scholarly record. Provision of the scholarly record for current and future generations is the primary mission of these research libraries. (Courant 2010)

This statement highlights the needs of researchers and scholars to discover, access, and cite the scholarly record over time, and of libraries to preserve the scholarly record and enable these activities. Something that immediately stands out about these needs is the expansive scope. The needs are not to preserve works from a particular country or time, by a particular author or set of authors, or in a particular format or medium (for instance print or analogue versus digital, or even born digital). The needs relate to all materials that can be used to further scholarship. HathiTrust acknowledges the scope of these needs in its broad mission “To contribute to the common good by collecting, organizing, preserving, and communicating the record of human knowledge” (HathiTrust n.d.a.). It acknowledges these needs also, and some particular points of strategy in addressing them, in its initial goals. One of most important points of strategy is that the effort to address the needs of HathiTrust’s designated community should and must be collective: “co-owned and managed” as the goals state, by the institutions ultimately responsible for the digital archive. The goals are as follows:

- To build a reliable and increasingly comprehensive digital archive of library materials converted from print that is co-owned and managed by a number of academic institutions.
- To dramatically improve access to these materials in ways that, first and foremost, meet the needs of the co-owning institutions.
- To help preserve these important human records by creating reliable and accessible electronic representations.
- To stimulate redoubled efforts to coordinate shared storage strategies among libraries, thus reducing long-term capital and operating costs of libraries associated with the storage and care of print collections.

- To create and sustain this “public good” in a way that mitigates the problem of free-riders.
- To create a technical framework that is simultaneously responsive to members through the centralized creation of functionality and sufficiently open to the creation of tools and services not created by the central organization. (HathiTrust n.d.a)

The initial goals center broadly around collections and collaboration: assembling a digital collection of materials, as comprehensive as possible; providing access to the materials; preserving them; and then using the materials in broader strategies that benefit first the partners, and by extension a larger worldwide community. Libraries’ goals to improve access to materials (including discovery and use) require their willingness and ability to address a number of specific micro-level challenges such as proper identification, description, and rights determination of materials (gaining a knowledge of what we have in our collections). Understanding what we have in our collections, the relationships between individual items, and the items’ rights statuses are precursors to the development of individual and collective strategies for macro-level challenges such as managing print and digital collections, expanding of lawful uses of in-copyright materials, and in general, improving our collective preservation infrastructure. The common characteristic of all of these challenges and strategies is that they are *big*. They lend themselves to, and can be best responded to by, collective action, at scale. Partnering Institutions support HathiTrust specifically as a platform to address their needs in these areas and facilitate this kind of collective action.

4.2 Preservation and Access

HathiTrust is a “light” archive and as such, strives to provide as much access as legally possible to all materials in the repository. Works that are determined to be in the public domain, or that rights holders have opened access to, are available to be read online as well as downloaded, subject to third party agreements.³ HathiTrust recognizes legal constraints and contractual obligations on materials, but does not preserve materials that depositors would wish to be stored without access, when access might otherwise be lawfully granted. For works that are in-copyright, HathiTrust includes full-text OCR in its repository-wide full-text search index, so that even though they are not available for reading or download, in-copyright works can be searched to retrieve word or query frequencies that may assist in determining the relevancy of a work or in locating specific information in a hard-copy of the work.

HathiTrust’s “light” orientation benefits users, as it provides access to a tremendous body of materials. It also has benefits for preservation, as the processes of retrieving and displaying data provide an additional check on the integrity of objects, and access in general gives the digital objects the best chance to be used and valued in the community, and therefore preserved into the future. The goal of providing access to preserved works manifests itself in numerous ways throughout HathiTrust’s technological infrastructure.

4.3 Openness

The last of HathiTrust’s goals speaks to a technical framework that provides significant centralized functionality, but is also open to distributed development of tools and services. This orientation and

³ Full download of materials is available where no restrictions exist. In most cases Google-digitized materials, which make up the largest group of materials where restrictions exist, are only fully downloadable by members of HathiTrust partner institutions.

general strategy towards openness extends to all aspects of the repository, from content formats to hardware and software, to organizational structure. The general strategy is that the long-term sustainability of the repository is served to the degree to which it is possible for member institutions to make use of the collective assets and services of the partnership, and to contribute to and manage them as well. The impact of HathiTrust's strong commitment to openness will be discussed further below.

5. Technical Infrastructure, Social System

HathiTrust's technical infrastructure was designed to meet the needs of its designated community, which can be categorized broadly on one hand as reliable long-term access to materials, and on the other as more efficient management of materials and resources to this end. This needs-based approach has resulted in a step-wise, modular trajectory to repository development, where discrete components that fulfill the needs for preservation and access interoperate as an integrated whole.⁴ It has also resulted in very practical decisions about these components that are fully cognizant of the concerns (including economic, technological, and sociological) of the designated community. The ways that HathiTrust has addressed core challenges in digital preservation, as articulated by the Task Force on Archiving of Digital Information's report in 1996 (which, as it has been noted, was a significant force in the development of TRAC and was used in the development of the OAIS model) is given below. It will be helpful before entering into a discussion of these elements to give a general description of the architecture and design of the repository.

5.1 Repository

The HathiTrust repository was developed according to the framework for Open Archival Information Systems and the Trustworthy Repository Audit and Certification criteria. The overall considerations for operation at scale, preservation and access, and openness have resulted in a strong drive for consistency and standardization across the repository. Consistency and standardization facilitate the operation of generalized processes across the repository for purposes of ingest and preservation (e.g., content auditing and reporting, replication, backup), as well as access (e.g., full-text search indexing, access to users through a variety of interfaces, collection-building capabilities). The major components of the infrastructure, shown in Figure 1, include:

- Ingest: processes check the fixity of objects received for deposit, transform them to HathiTrust specifications if needed, perform rigorous validation, package objects for ingest, and finally bring them into the repository.
- Archival Storage: HathiTrust storage consists of two geographically separated instances of the repository on spinning disk, with tape backup stored in a third location.
- Data Management: HathiTrust manages bibliographic and rights information about objects, as well as information about the print holdings of partner institutions that correspond with HathiTrust's digital holdings. The significance of managing partner print holdings will be discussed further below.

⁴ See York 2010 for a detailed discussion of repository architecture.

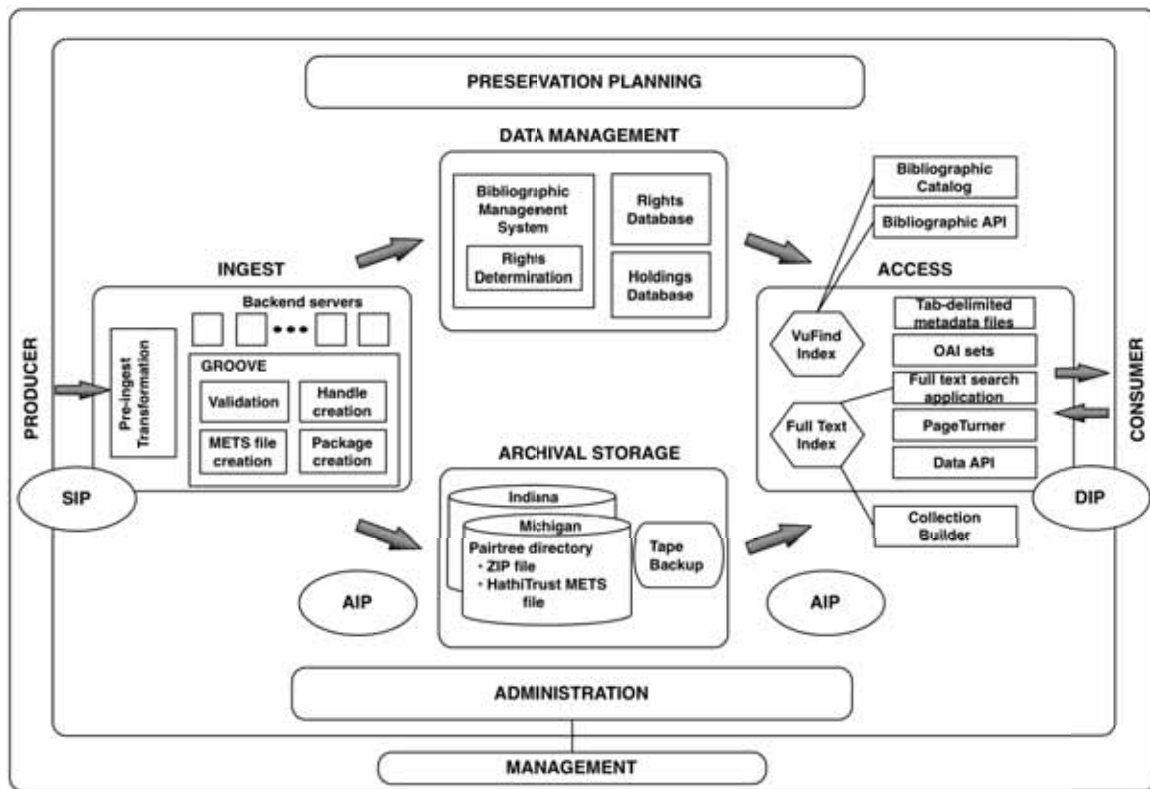


Figure 1. HathiTrust architecture according to the OAIS model (York 2010).

- Access: access services include
 - Bibliographic and full-text search of all materials
 - Reading and download capabilities for public domain and open access materials
 - The ability to assemble virtual collections of materials (i.e., “book bag” functionality)
 - A variety of APIs and data feeds for both bibliographic data and repository content (images and OCR text)

5.2 Preserving Digital Information

In describing the overall landscape of digital information and preservation, the Task Force on Archiving of Digital Information notes that:

The process of preserving digital information will vary significantly with the different kinds of objects – textual, numeric, image, video, sound, multimedia, simulation, and so on – being preserved. Whatever preservation method is applied, however, the central goal must be to preserve information integrity; that is, to define and preserve those features of an information object that distinguish it as a whole and singular work. In the digital environment, the features that determine information integrity and deserve special attention for archival purposes include the following: content, fixity, reference, provenance, and context. (1996, 13)

Each of these elements, content, fixity, reference, provenance, and context, will be taken in turn.

5.2.1 Content

With regard to content, the Task Force states:

The measure of integrity in the preservation process thus turns, at least in part, on informed and skillful judgments about the appropriate definition of the content of an digital information object—about the extent to which content depends on its configuration of bits, on the structure and format of its representation, and on the ideas it contains—and for what purposes. (1996, 13)

At the broadest level, considerations about content in HathiTrust begin with what is selected for digitization and preservation. The materials ingested by HathiTrust to-date have been those digitized and submitted individually by the partnering institutions (i.e., materials determined by those institutions to be of enduring value). HathiTrust formed a Collections Committee in 2010, however, whose charge includes making recommendations about content in HathiTrust (HathiTrust n.d.b), and partners recently approved a targeted initiative surrounding United States federal government documents (HathiTrust n.d.c). These initiatives underscore opportunities for collective decision-making about the addition of materials to the repository in the future.

Apart from selection of materials at an intellectual level, HathiTrust has defined parameters for the general types of materials as well as the content formats and specifications that are accepted. The general types of materials HathiTrust preserves at a production level currently are digitized books, journals, and book-like materials, such as codex manuscripts. Pilot projects involving image (e.g., maps and photographs), audio, and born-digital content are underway. For the books and book-like materials, there are only three formats in the repository that are primary targets of preservation: ITU G4 (bitonal) TIFF images, JP2 images, and Unicode text (HathiTrust volume packages include both plain text Optical Character Recognition (OCR) text and OCR with word coordinate location information). HathiTrust enforces adherence to these formats (including validity), minimum resolution standards, and internal image metadata specifications through rigorous validation processes on ingest. The specific types and numbers of formats in HathiTrust are not important in and of themselves (HathiTrust will undoubtedly support more formats and types of materials at a production scale over time), but are important the degree to which they satisfy a variety of community concerns. For example, ITU G4 TIFF, JP2 and Unicode are standard and open formats that meet community-accepted standards for digital preservation. They are also widely supported on a number of platforms and not dependent on particular hardware or software to render to users. These attributes of the formats inspire confidence in the community in their ability to be preserved and migrated forward to new preservation formats over time.

The formats HathiTrust accepts express its orientation toward openness, which in this case facilitates preservation of materials. The openness of formats facilitates access to materials as well, however, and access in particular at scale. The openness and flexibility of the formats allows them to be transformed on the fly to formats that can easily be downloaded or displayed to users on the Web. Management of files is thus simplified, as derivative images do not need to be stored in the repository, and repository systems do not need to be developed to maintain and disseminate them. The openness of the formats allows a uniformity of content across the repository that lowers the overhead of cross-repository management functions while offering a variety of access options to users.

It is relevant to note that HathiTrust has benefitted greatly from the uniformity of Google digitization with regard to content format and standards. The fact that millions of volumes have been digitized in the same way to the same specifications has greatly facilitated rapid growth of the repository (HathiTrust has

grown overall from 2.5 million volumes from its launch in 2008 to 10.5 million in 2012). HathiTrust has encountered challenges when seeking to accept content digitized from other sources, even when those sources use the same formats. This is due to the issues and work involved in transforming content to meet HathiTrust specifications, including assembling content and metadata into HathiTrust content packages. The collective work of the HathiTrust community has been key in addressing this challenge.

HathiTrust partners have worked closely together to define specifications and process for transforming content from large-scale digitization sources such as the Internet Archive, and develop tools, available from the HathiTrust website (HathiTrust n.d.d), that allow partners to transform, validate, and package content that is digitized on a smaller scale to HathiTrust specifications prior to submission. Working collaboratively, HathiTrust institutions have developed a framework that allows institutions to participate deeply in the preservation of their content, while lowering the overall costs to the repository of staging and transforming content, some of the highest costs associated with digital preservation repositories.

5.2.2 Fixity

Fixity in the Task Force report refers to the way content is “fixed as a discrete object,” with the concern that objects might be changed or corrupted without notice (1996, 14). The concept of fixity relates closely to the concept of authenticity, as articulated by Luciana Duranti (Duranti 1995) and authenticity and integrity, as discussed by Clifford Lynch (2000). These relationships will be explored more closely below.

HathiTrust verifies the fixity of objects internally at several levels. The first is through verification, when possible,⁵ of checksums for content as a part of the ingest process (calculating a message digest for content in the Submission Information Package and comparing it with the digest provided with the content). The second is by periodically re-calculating the checksums of objects in the repository and comparing them with checksums generated prior to ingest.⁶ The third is through data integrity mechanisms internal to the storage itself, which use checksums to ensure that data transferred from one storage site to another are not corrupted, and to detect and automatically repair errors, including those caused by “bit rot” phenomena such as misdirected or torn writes.

HathiTrust communicates fixity, to a degree, to users as well, through the use of watermarks on images displayed in or downloaded from HathiTrust Web interfaces. The watermarks are not actually inscribed into the images themselves; they are overlaid on derivative images when the derivatives are created from the master files. It is thus possible to tell from the Web and printed copies that the images came from HathiTrust, although the watermarks do not have meaning for internal tracking.

These represent some of the mechanisms HathiTrust has in place. As Clifford Lynch has discussed, however, issues of authenticity and integrity at their base are largely functions of trust and context (2000). In discussing checks of internal consistency using checksums that are calculated for objects, he notes that when such a checksum or digest is used, “our confidence in the integrity of the object is only as good as our confidence in the authenticity and integrity of the digest.” In such a situation, the link between the claim that a message digest is correct and the claim that an object maintains its integrity “is done by association and context—by keeping the claim bound with the object, perhaps within the scope of a

⁵ It is possible that materials desired for ingest are not accompanied by valid checksum information.

⁶ Checksums are recorded in metadata that is stored with objects in the repository.

trusted processing system such as an object repository.” Put in other words, it is within a trusted environment that claims of authenticity and integrity have their meaning.

HathiTrust uses automated checks on integrity to detect random or accidental corruption of objects in the repository, but these mechanisms would likely not be sufficient to ensure integrity in the event of a successful intentional attempt to corrupt content. HathiTrust’s multiple levels of redundancy (multiple storage locations and backup) could be used to restore any or all objects in the repository following such an act. A key point regarding fixity, however, is that it is broader mechanisms of system security (to prevent malicious forces from outside) and trust in staff (to ensure security from inside) that ensure the integrity of the overall environment, and give validity to further internal checks that are performed. As the fixity and integrity of objects depends to a significant degree on trust in the people and social system (the libraries) operating the repository, it is essential for the libraries participating in HathiTrust to maintain the trust of the community in this social system, as well as and including the technical systems it has in place.

5.2.3 Reference

With regard to reference, the Task Force states: “For an object to maintain its integrity, its wholeness and singularity, one must be able to locate it definitively and reliably over time among other objects” (1996, 15). HathiTrust addresses issues of reference in several ways. The first is the way items in the repository are identified. When an object enters the repository it is assigned an identifier that is composed of the identifier for the object prior to when it entered the repository, if available, and a namespace. HathiTrust prefers to use identifiers for objects that are in use by the depositor (in the case of digitized books this is often the barcode of the physical volume) if they have good identifier qualities, including guaranteed uniqueness (HathiTrust n.d.e). This is to avoid maintenance that would be involved in mapping and updating HathiTrust-generated identifiers, and to facilitate references by institutions to representations of their materials in HathiTrust. Namespaces are selected by the depositor and are used to identify the depositing source, as well as distinct identifier schemes of submitted objects. If items from a depositor have more than one identifier scheme, more than one namespace is used (HathiTrust n.d.e). As an example, the University of California uses the namespaces “uc1” and “uc2” to distinguish volumes digitized by Google and by the Internet Archive, each of which have distinct identifiers schemes. An example of an identifier in each group is given below:

uc1.b3543486 (Google-digitized)
uc2.ark:/13960/t26973133 (Internet Archive-digitized)⁷

HathiTrust thus takes great care to ensure the unique identification of items in the repository and enable references to original items where possible.

HathiTrust further enables reference through the structure of the repository. The objects in HathiTrust are stored in directories in one large file system. The repository uses a Pairtree structure, which maps identifier strings to directory paths for digital objects pair-wise, with the name of the final directory being the object identifier (Kunze et al. 2008). For example, the path on the repository file system to the directory of the item “uc1.b3543486” is ../uc1/pairtree_root/b3/54/34/86. The files of the object itself are located in this directory and named b3454386.zip and b3454386.mets.xml. The zip file contains the content files of an object—the images and OCR for digitized books—as well as additional

⁷ These are the same examples used in York 2010.

content metadata. The XML file contains a variety of technical, administrative, and structural metadata encoded in the Metadata Transmission and Encoding Standard (METS) (Library of Congress n.d.a), that serve purposes both of preservation and access.⁸ There are several benefits to using the Pairtree structure, including that it a) ensures that objects are uniformly accessible to repository systems and access services; b) makes it easy for content to be imported, understood, and used in new storage system without the system knowing anything about the nature or contents of the stored objects; and c) allows object operations such as backup and restore, to be performed using native operating system tools, facilitating disaster recovery (Kunze et al. 2008). These benefits have clear advantages for reference, as objects can be located and operated on through automatic processes. They also inherently facilitate operation at scale, preservation and access, and openness—openness to the degree that HathiTrust objects are not tied to the specific infrastructure they are stored on and could either be operated on by tools independent of the software used in current storage, or moved to totally different storage and operated on immediately.

There are three further ways that HathiTrust facilitates reference. The first is by embedding the identifier of objects in the metadata of images that make up digital volumes themselves.⁹ The second is through the creation of unique and permanent identifiers for objects using the Handle System (Corporation for National Research Initiatives, n.d.). Permanent identifiers comprise a Handle namespace and the HathiTrust identifier, which are combined together to form a permanent URL where the object can be located on the Web. HathiTrust also provides the date of the most recent version of the volume in HathiTrust, facilitating citation (versioning is discussed in the section on Provenance below).

As in other areas, the mechanisms that HathiTrust uses to facilitate reference depend on broader social factors—for instance, the selection and use of identifiers by depositing institutions; the reliability of the Handle service. By taking a stance that is sensitive to these factors (e.g., favoring the use of existing identifiers, using a uniform scheme of structure and reference in the repository), HathiTrust positions itself to be responsive to them and as needs for and applications of reference capabilities change over time.

5.2.4 Provenance

The Task Force report highlights two ways that establishing the provenance of objects serves to preserve their integrity:

First, a tracing of chain of custody from the point of creation helps to create the presumption that an object is authentic, that it is what it purports to be and that its content, however defined, has not been manipulated, altered or falsified (Duranti 1995: 7-8). The second effect of establishing provenance through a chain of custody is to document, at least in part, the particular uses of the object by the custodians. (1996, 17)

HathiTrust traces the provenance of digital objects by recording the original source of the material represented in HathiTrust, the agent of digitization, and a variety of administrative (including provenance and preservation) metadata about objects, where this metadata is available.¹⁰ HathiTrust uses the

⁸ Details about HathiTrust content packages, and the metadata contained in the XML file are available at (HathiTrust n.d.f). See also York 2010.

⁹ In the `DocumentName` element of TIFF files and the `dc:source` element of JP2 files.

¹⁰ At a minimum, HathiTrust requires information about the digital capture of items. An explanation is needed if this information is not available. HathiTrust accepts other information relevant to provenance and preservation of materials prior to their entry into HathiTrust, but does not require it.

Preservation Metadata: Implementation Strategies (PREMIS) standard (Library of Congress, n.d.b) to record preservation metadata, including the time and date of digital capture, transformations that may have been performed, fixity checks, validation, quality review, and other events that may occur either prior to HathiTrust taking responsibility of a digital object or subsequently.

HathiTrust does not keep multiple versions of objects in the repository.¹¹ If a new version of an object is available for ingest, as is often the case in particular with Google-digitized volumes,¹² the new version overwrites the existing version. A new series of preservation events (e.g., fixity check, digest calculation, validation, and ingest) is written into the PREMIS metadata for the re-ingested object, and all previous events are retained, providing a means to determine whether and how many times an object has been ingested. HathiTrust uses this practice primarily because of the immense scale of the repository and the frequency with which volumes from Google are reprocessed and made newly available. The cost of preserving multiple versions of Google-digitized volumes would drastically increase the cost of preservation to partners, and the value of retaining these versions is not clear. For instance, it is important for users to be able to reliably cite the version of an item that they used (a version date is provided for this purpose as mentioned above). It is also important for HathiTrust to record changes that are made to volumes where possible.¹³ Whether or not it is important to record Google's or another entity's attempts to create a reliable representation of a known object is a separate question. The Task Force report acknowledges the relationship between documenting provenance and the concepts of fixity and authenticity, and factors affecting and complicating notions of authenticity have been discussed in the section on Fixity above. It is worth exploring these factors a little more deeply, however, particularly in relation to the concept of reliability, as it relates to issues of versioning and is raised in the article by Duranti that the Task Force cites.

A primary point that Duranti makes in the article is that the “concepts of authenticity and reliability must be kept intellectually separate”¹⁴ A danger in conflating the two, or of focusing primarily on the integrity and authenticity of records, is that records may be completely authentic, but this does not mean that they are reliable (1995, 7). Duranti states that, “A record is considered reliable when it can be treated as a fact in and of itself, that is, as the entity of which it is evidence” (1995, 6). According to Duranti, the elements that provide a record with reliability are its form and its procedure of creation (“the body of rules according to which acts or portions of them are recorded”) (1995, 6). Duranti notes, “A record is regarded as reliable when its form is complete, that is, when it possesses all the elements that are required by the socio-juridical system in which the record is created for it to be able to generate consequences recognized by the system itself.” Some elements that commonly contribute to form are signature and date

¹¹ Versions here refer to multiple different copies of a distinctly identified object (for instance, one version of an object that is missing a page and a corrected version that is not). HathiTrust does preserve multiple editions of the same work, as well as multiple copies of the same work that are in the repository under different identifiers (i.e., duplicate volumes).

¹² Google is constantly improving the algorithms it uses to process the raw images it captures of library volumes. When new versions of volumes from any institution are available, if they pass a certain quality threshold, they are re-ingested into HathiTrust.

¹³ The partners have observed a trend in higher quality scans returned from Google over time, but an automated mechanism to determine whether or how much the quality of a given volume has changed following reprocessing by Google (and correspondingly, what specifically has changed) does not exist. This can only be determined through manual inspection, though many minute changes would likely not be detectable.

¹⁴ (Duranti 1995, 8). Duranti's concern is primarily with electronic records, but many of the same concerns apply to digitized volumes.

of creation. Procedures of creation might include appropriate responsibility for signing, recording of facts by multiple persons, or distribution to multiple addresses (1995, 6). Duranti states that the same elements, completeness of form and procedure of creation, determine the reliability of copies that are made of originals, and acknowledges that there are different degrees of reliability, ranging from a simple copy made “without the dating and attestation of the copying person,” to copies that might be more reliable than the originals themselves (1995, 7).

The digitized volumes found in HathiTrust are intended to be copies of the original physical volumes.¹⁵ However, while HathiTrust records the date images of original volumes were created, in most cases (except when files are received directly from the publisher) there is no official entity that verifies the reliability of the copies with respect to the way they represent the originals. In the socio-juridical context of library digitization, the reliability of digital copies, their ability to stand for the items they represent and to generate consequences (i.e., be cited and used as surrogates for the physical volumes), has generally been established by libraries through quality assurance. Whether libraries conduct the digitization of items themselves or contract with a vendor, there is an underlying assumption, because of the trust society places in libraries and libraries’ living up to this trust over time, that appropriate steps have been taken to ensure that digitized copies accurately represent the original items, or that steps can be taken to address problems that are encountered.

In this context, where libraries have primary responsibility to their communities for ensuring the reliability of digitized items they make available, the question of retaining versions of items, and broad questions of reliability in general, become questions that depend on the needs and resources of the social system. The question in this case becomes not whether to retain multiple versions of items, but how to best meet community needs for preservation, including reliability and authenticity, in ways that are sustainable and responsibly manage community resources. Recording the provenance of digitized items is crucially important, but does not in and of itself speak to issues of reliability, and can be completely separate if a digitized item has quality problems and is not able to stand as a faithful copy of the original it is intended to represent. Issues of reliability, particularly in relation to information quality and in light of community needs, are a current area of study for the partnership.¹⁶

5.2.5 Context

Content in the Task Force report refers to “the ways in which [digital information objects] interact with elements in the wider digital environment” (1996, 18). The report points to three dimensions of context that have to do with technical aspects (hardware and software dependencies), linkages among digital objects (to the degree to which the integrity of an object lies in the network of linkages), and communication medium (the extent to which the way materials are distributed—for instance, bandwidth or security constraints or attributes—account for characteristics of the digital objects) (1996, 18-19).¹⁷ The

¹⁵ HathiTrust’s Digital Preservation Policy notes that “HathiTrust is committed to preserving the intellectual content and in many cases the exact appearance of materials that have been digitized for deposit.” This includes “Digital representations (images) of content as the content appeared in its original form, with the same layout and colour (e.g., for illustrations and artwork), and in the same order.” (HathiTrust n.d.g).

¹⁶ See, for example, the work of Paul Conway (Conway 2011). HathiTrust’s policy on quality is available at <http://www.hathitrust.org/quality>.

¹⁷ With regard to communication medium, the report gives the example of increasing bandwidth resulting in the production and distribution of high-bandwidth products such as “full-motion video.”

report points also to a broader social environment and the contextual role that policies and implementation details regarding bandwidth, security, and other network qualities can have on information integrity.

HathiTrust's use of open formats and practice of transforming master images for access in different contexts address many of these contextual concerns about information integrity (the integrity of objects in HathiTrust does not depend on hardware or software, and due to the flexibility of formats, access considerations are separate to some degree from preservation concerns). HathiTrust objects exist entirely within the repository, so the issue of linkages as it is relayed does not apply. There are some significant elements of context in HathiTrust, however, that are relevant beyond the explicitly digital environment in which they exist. The first of these is the relation of objects in HathiTrust to their print counterparts, and the second has to do with discovery and use.

Relation to Print. HathiTrust's fourth stated goal is "To stimulate redoubled efforts to coordinate shared storage strategies among libraries, thus reducing long-term capital and operating costs of libraries associated with the storage and care of print collections." Understanding the relation of the digital objects in HathiTrust to the print items owned by libraries (whether the item in HathiTrust was digitized by their library or not) has had profound implications for the development of HathiTrust. HathiTrust began with a pricing model based on a per-gigabyte fee, covering the infrastructure costs of the content institutions deposited. As HathiTrust has grown (from 2.5 million volumes when it was launched to nearly 10.5 million today), the overlap with North American academic and research libraries has become so significant—likely more than 50% on average with Association of Research Library libraries¹⁸—that it has shifted from being a strategy for institutions to preserve their digital volumes, to being a strategy to preserve their print volumes as well (with digital "backups").

In recognition of this fact, HathiTrust developed a pricing model, which will be in effect in 2013, that is based on the overlap of partnering institutions' print holdings with the digital holdings in HathiTrust. The pricing model is supported by a holdings database (represented in the Data Management component in Figure 1) that maps institutional print holdings to holdings in HathiTrust. This holdings database represents a new contextualization of the digital objects, and the physical objects as well, in a broader information environment. In addition to helping libraries to understand the relationships between their collections of print and digital objects, the holdings database will support the expansion of lawful uses of in-copyright materials in HathiTrust that are owned in print by the partnering libraries.¹⁹ Contextualizing the holdings of HathiTrust revolutionizes the way libraries conceive of their collections and provides a basis for a "deep infrastructure" on which libraries can collaboratively move their content and services into the future.

Discovery and Use. HathiTrust offers centralized access services, including bibliographic and full-text search, as well as reading, downloading, and collection-building capabilities. In addition to these services, and in support of the last of its stated goals related to a centralized yet open technical framework, HathiTrust offers several APIs and data feeds that allow partner and non-partner institutions to contextualize their own collections in relation to HathiTrust. For instance, partners can use information from HathiTrust APIs to add links or entire records for HathiTrust items to local discovery mechanisms.²⁰

¹⁸ This figure is extrapolated based on analysis of trends observed in (Malpas 2011) and HathiTrust repository growth since that time.

¹⁹ A description of the specific uses and circumstances of access to in-copyright works that HathiTrust has targeted is given at http://www.hathitrust.org/authors_guild_lawsuit_information#Details.

²⁰ Information on how this can be done is available at (HathiTrust, n.d.h).

The collection-building capability allows users to create and reference canonical bibliographies of materials such as the English Short Title Catalog, and other sets of materials.²¹

In each of these ways, by contextualizing HathiTrust materials in their broader environment, and allowing others to contextualize local collections, HathiTrust creates pathways for meaningful connections between collections and items to be made. This simultaneously improves discovery and use of materials, and uncovers new opportunities for libraries and cultural heritage institutions to engage in collective action to address shared challenges.

6. Conclusion

This paper has highlighted the ways that HathiTrust's understanding of preservation as a social and collaborative activity has influenced specific approaches it has taken to preserving digital information—both in its technical infrastructure, and in relation to issues of content formats, fixity, reference, provenance, and context. By focusing on community needs and social factors in concert with technical considerations, HathiTrust has been able to gain a broad base of support for its activities, and take decisions that strengthen its ability to meet community needs over the long-term. In the end, it is we, collaboratively, who have responsibility for moving our collections into the future, and strategies that bring our efforts closer in concert with one another are the most likely to succeed.

Acknowledgements

This paper has only been possible through deep consultation and collaboration over a period of years with staff at the University of Michigan Library that designed and built HathiTrust's repository infrastructure, and staff at numerous institutions as the partnership has grown. I would like to extend special thanks to Cory Snavely, Aaron Elkiss and Sebastien Korner for their input and clarification of repository processes, and John Wilkin for reviewing and commenting on the paper.

References

- Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). Washington, D.C: CCSDS Secretariat, 2002; Program Integration Division (Code M-3), National Aeronautics and Space Administration.
- Conway, Paul. "Archival Quality and Long-term Preservation: a Research Framework for Validating the Usefulness of Digital Surrogates." *Archival Science* 11, no. 3 (2011): 293–309.
doi:10.1007/s10502-011-9155-0.
<http://www.springerlink.com/content/9322j77m6327h123/abstract/>.
- Corporation for National Research Initiatives. "Handle System." <http://handle.net/>.
- Courant, Paul. "Testimony of Dean Paul Courant." Fairness Hearing on Proposed Settlement, February 18, 2010. <http://www.lib.umich.edu/michigan-digitization-project/fairness-hearing-testimony-of-dean-paul-courant>.

²¹ The collection of English Short Title Catalog materials is available at <http://babel.hathitrust.org/cgi/mb?a=listis;c=247770968>. See also HathiTrust's featured collections and the full list of user-created public collections: <http://babel.hathitrust.org/cgi/mb?a=listes#all>.

- CRL, and OCLC. Trustworthy Repositories Audit and Certification: Criteria and Checklist. 2007. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf.
- Duranti, Luciana. "Reliability and Authenticity: The Concepts and Their Implications." *Archivaria* 1, no. 39 (1995): 5-10. <http://journals.sfu.ca/archivar/index.php/archivaria/article/view/12063/13035>.
- HathiTrust. n.d.a. "Mission and Goals | HathiTrust Digital Library." http://www.hathitrust.org/mission_goals.
- . n.d.b. "Collections Committee Charge | HathiTrust Digital Library." http://www.hathitrust.org/wg_collections_charge.
- . n.d.c. "Constitutional Convention Ballot Proposals | HathiTrust Digital Library." http://www.hathitrust.org/constitutional_convention2011_ballot_proposals#proposal4.
- . n.d.d. "Ingest Tools | HathiTrust Digital Library." http://www.hathitrust.org/ingest_tools.
- . n.d.e. "Guidelines for Digital Object Deposit | HathiTrust Digital Library." http://www.hathitrust.org/deposit_guidelines#pdi.
- . n.d.f. "Digital Object Specifications (METS and PREMIS) | HathiTrust Digital Library." http://www.hathitrust.org/digital_object_specifications.
- . n.d.g. "Digital Preservation Policy | HathiTrust Digital Library." <http://www.hathitrust.org/preservation>.
- . n.d.h. "Data Availability and APIs | HathiTrust Digital Library." <http://www.hathitrust.org/data>.
- Kunze, John, Martin Haye, Erik Hetzner, Mark Reyes, and Cory Snavelly. "Pairtrees for Object Storage." 2008. <https://confluence.ucop.edu/download/attachments/14254128/PairtreeSpec.pdf?version=1>.
- Library of Congress. n.d.a. "Metadata Encoding and Transmission Standard." <http://www.loc.gov/standards/mets/>.
- . n.d.b. "PREMIS: Preservation Metadata Maintenance Activity." <http://www.loc.gov/standards/premis/>.
- Lynch, Clifford A. 2000. "Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust." In *Authenticity in a Digital Environment*. CLIR Papers. Washington, D.C: Council on Library and Information Resources. <http://www.clir.org/pubs/reports/pub92/lynch.html>.
- Malpas, Constance. "Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment." OCLC, 2011. <http://www.oclc.org/resources/research/publications/library/2011/2011-01.pdf>.
- Research Libraries Group, and OCLC. *Trusted Digital Repositories: Attributes and Responsibilities*. Mountain View, CA: Research Libraries Group, 2002. <http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf>.
- Task Force on Archiving of Digital Information. "Preserving Digital Information : Report of the Task Force on Archiving of Digital Information." Commission on Preservation and Access and Research Libraries Group, 1996. <http://library.oclc.org/cdm/singleitem/collection/p267701coll33/id/272/rec/11>.
- York, Jeremy. "Building a Future by Preserving Our Past: The Preservation Infrastructure of HathiTrust Digital Library." In session 157 - ICADS with Information Technology. Gothenburg, Sweden, 2010. <http://www.hathitrust.org/documents/hathitrust-ifla-201008.pdf>.

"Archiving-as-a-Service"

Influence of Cloud Computing on the Archival Theory and Practice

Hrvoje Stančić,¹ Arian Rajh² and Ivor Milošević³

¹University of Zagreb, Croatia, hrvoje.stancic@zg.t-com.hr; ²Croatian Agency for Medicinal Products and Medical Devices, arian.rajh@halmed.hr; ³University of Zagreb, Croatia, ivor@srce.hr

Abstract

The research is positioned in the context of the responsibility of archives to preserve important records in an increasingly changing technological environment, and focused on the impact of cloud solutions on archival theory and practice. Authors address several questions which they consider crucial for archival science and community. Results of the survey on the usage of private cloud are given. In view of that, the authors examine if the concept "Archiving-as-a-Service" will require redefinition of archival practice in the new technological and organizational context. Finally, they suggest the need for transition from postcustodial to "postcustodial 2.0" paradigm.

Authors

Dr. Hrvoje Stančić, associate professor, is Head of the Chair for archival and documentation science at the Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia with national and international teaching involvement. He is author of the book 'Digitization' and more than 50 scientific papers. His research is focused on digitization and long-term digital preservation issues. He co-coordinated activities of DigitalPreservationEurope initiative at the Faculty level. He was researcher at several national and European research projects and was the leader for national component of the EU 'Heritage Live' cross-border research project.

Dr. Arian Rajh, assistant professor, works as Head of document, records and project management department in Croatian Agency for Medicinal Products and Medical Devices. He is also a member of European Telematic Implementation Group on Electronic Submission, chairperson of Ad-hoc drafting guidance subgroup 'Best archiving practice' and external associate at Faculty of Humanities and Social Sciences, University of Zagreb, Croatia. He graduated in Information sciences and Comparative literature in 2005 and received his Ph.D. in archival science in 2010. Rewards/scholarships: City of Zagreb scholarship (2004), Rectorial award (2004), "Franjo Marković" (2005), DigitalPreservationEurope Exchange Programme Award (2008).

Ivor Milošević, is a Ph.D. student at the Doctoral study of Information and Communication Sciences at the Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia. He works at the University Computing Center where he is responsible for implementation of virtualization services.

1. Introduction

Paradigm is the leading form of a specific science and its institutional position and practice in a certain period, where explicit ideas became dominant theoretical framework. Other theories come to be included in the newly-born paradigm. Then the paradigm-creation period is followed by formation of "normal science" and its practice which influences various institutional practices. However, and this is not in line with Kuhn's model of development of natural and technical sciences, paradigms in humanities and social

sciences can be seen as marketplaces of ideas or islands of coexistence of similar ideas.¹ Drastic cuts of paradigms are less often noticed in models of development of humanities and social sciences, theories pertaining to these domains can be cumulative and knowledge can be built on previously known knowledge—if the differences are not radical. Differences between paradigms in archival science are more distinct than in social sciences in general, but they do not have all characteristics of scientific revolution model. Paradigms in archival science that can be noticed, at least in its scientific period, can be described as codification of archival science (Dutch archivists and Jenkinson), modern archival science (Schellenberg and followers) and postmodernism or, as John Ridener² stated, the “questioning” paradigm. Principle and observation of changing paradigms can be applied also to archival institutions and their activities and interactions with the creators, and good example of that is the difference between custodial and post-custodial archives. Notion of post-custody, from Gerald Ham onwards, from distributed custody methods to broader connotation of the term that includes new functions of archival institutions and services are related to the latest archival paradigm.

Criteria for identification of a possible new archival paradigm is the difference between the basic theoretical framework (with which we can tie the largest number of archival theories in certain period, theories that significantly affect the practice and methodology, like, for example, emergence of other forms of archival description) and its practical results. Development of archival science is not fully in accordance with the scientific revolution model, as it could be expected, since archival science pertains to social sciences. Thus, the phenomenon of accumulation of knowledge, together with its reinterpretation and reuse can be detected. However, professionally recognized body of theories grouped under the same paradigmatic umbrella is different comparing to the previously known theories. There are paradigmatic differences in basic definitions, like definitions of archives, the role and importance of professionals in the archives, and their basic methodological principles. Postmodern archival theorists raised Derridian question of what is external to archive but organizes its practice. Today’s question is similar—we may ask ourselves what is external to archive but aims to influence its practice. This is the reason to start reflecting on today’s and future services that can support archival functions. That is why the notion of archiving and cloud computing should be seen in the light of the latest paradigm, even as its contemporary enhancement like “post-custody 2.0” or similar.

Records continuum model places the object of archival preservation in a specific temporal and spatial relationships of the primary and any other institutional, business, or social context in order to describe broader preservation responsibilities. Can cloud services and archiving organized as cloud services be considered as means of fulfilling those responsibilities? Can they be “mapped” to dimensions and axis of the records continuum model? Cloud services can support organization needs of creators and institutions with limited and uncertain IT resources and thus they can facilitate better availability of records and archives to the future users, assuming good management of the service. Cloud service can also support recordkeeping function of different but linked creators, group their archives, document their wider purpose, and facilitate creation of collective memory. Today’s archival experts are active members of the information community so they should actively think about new storage and (potential) archiving possibilities and plan their responses to the challenges. And challenges are approaching quickly due to the quick technological development. We are witnessing constant technological change and since now, with

¹ Thomas S. Kuhn, *The Structure of Scientific Revolutions*, 3rd ed. (Chicago: University of Chicago Press, 1996).

² John Ridener, *From Polders to Postmodernism: A Concise History of Archival Theory* (Duluth, MN.: Litwin Books, 2009).

the proliferation of cloud services, the postcustodial practice did not have a great challenge. However, it seems that the archival science will have to approach the archiving in the cloud issue.

2. Digital Preservation Environment

“With the current global economy facing financial pressure, organizations are compelled to reduce operational costs and streamline their efficiency. Responding to this imperative, it is estimated that more than 20 percent of organizations have already begun to selectively store their customer-sensitive data in a hybrid architecture that is a combined deployment of their on-premises solution with a private and/or public cloud provider in 2011. (...) At year-end 2016, more than 50 percent of Global 1,000 companies will have stored customer-sensitive data in the public cloud.”³ This Gartner’s prediction is showing a trend of business change towards the cloud-based storage. Since there are several approaches to provision of cloud services and organization of cloud storage it is necessary to distinguish between them in order to better understand their functionalities. This will help better understanding of necessity and possibilities of archival intervention.

The cloud computing paradigm (which can be regarded as part of the questioning paradigm or as “post-custody 2.0” in the context of archives and archival science) includes a spectrum of many models, developers and reference designs as well as several essential characteristics, as mandated by the National Institute for Standards and Technology, which include: on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service.⁴

In the standard classification of cloud infrastructures one differentiates between several deployment models with subsequent service models as follows:

- *Software as a Service (SaaS)*: ability to deliver applications from cloud-based physical infrastructure, accessible via various client software tools or devices. The user has no awareness or control of the underlying physical components or software configuration capabilities outside the delivered application.
- *Platform as a Service (PaaS)*: ability to deliver complete environments (operating systems and required tools) for testing or development of external applications. The user, however, has no control over the configuration settings of the application-hosting environment.
- *Infrastructure as a Service (IaaS)*: ability to deliver complete virtual datacenters to the user who is then able to configure and deploy virtual machines and other relevant/corresponding virtual components according to their personalized requirements.

According to the deployment models cloud implementations include:

- *Private cloud*: where it is implied that the cloud infrastructure is built and provisioned for private use by a single organization. Private clouds in practice tend to be service-oriented with specific roles and requirements.

³ “Gartner Reveals Top Predictions for IT Organizations and Users for 2012 and Beyond,” Gartner, December 1, 2011, <http://www.gartner.com/it/page.jsp?id=1862714>.

⁴ The NIST Cloud Computing Project. Accessed August 8, 2012, <http://csrc.nist.gov/nice/states/maryland/posters/cloud-computing.pdf>.

- *Community cloud*: where the physical infrastructure is implemented, administered, and operated by several organizations in a certain community of consumers from organizations that have shared goals and requirements.
- *Public cloud*: the cloud infrastructure is intended for “rent” by the public users, as delegated by the provider usually for profit or other means of compensation for the provider.
- *Hybrid cloud*: the combination of two or more physical cloud infrastructures from different branches of the above listed deployment models that are physically separate but are connected via the means of mutual data and application portability or management hierarchies.

Most initial, real world private cloud implementations tend to focus on either SaaS or IaaS methodology and features, or the combination of those two, since the role of PaaS is limited to a very specific set of users, mainly in software development where they deliver computing platform or a solution stack as a service, often consuming cloud infrastructure and sustaining cloud applications.

To adequately address increasing examples where combinations of these concepts are merged on both the physical and logical level in order to create a specific service (such as the proposed Archiving as a Service model) it is possible to resort to a form of reclassification of the standard paradigms and definitions in the cloud computing service environment, where these combinations are described and are given adequate alternate labels depending on the specific service they are ultimately intended to offer.

On the other hand, since most contemporary installations and implementations of both private and public cloud infrastructures resort to various underlying hardware technologies, software solutions as well as logical models, in order to create the all-encompassing concept of XaaS or “Everything as a service,” within which all other services and usage models can be compartmentalized, one can view the attempt to adequately describe archival service and digital preservation as a method of selecting required components and features within this base umbrella model.

It is possible to explain and define the concept of cloud computing as an additional security layer superimposed on virtualization technologies, allowing users a degree of self service and disassociating the underlying physical components from the end-user experience. So in the context of “digital preservation as a (cloud) service,” the cloud infrastructure would provide all the components which can be utilized towards the goal of enabling successful and efficient preservation of digital data.

Virtualization allows high availability schemes as well as safety mechanisms from data corruption due to hardware malfunction to be implemented. “The key here is replication of the storage system so that, in an emergency situation, the remote location can independently take over all the operating components of the primary.”⁵

It is certainly impossible to separate digital preservation solutions and particular technology completely, but the key for successful preservation is in using and switching to available technologies. The principle was put out in OAIS RM⁶ concept and standard—to observe technological environment and

⁵ Ivor Milošević and Hrvoje Stančić, “Usage of Virtualization Technologies in Long-Term Preservation of Integrity and Accessibility of Digital Data,” in *INFuture2011: Information Sciences and e-Society*, ed. Clive Billenness, Annette Hemera, Vladimir Mateljan, Mihaela Banek Zorica, Hrvoje Stančić, Sanja Seljan (Zagreb: Department of Information Sciences, Faculty of Humanities and Social Sciences, 2011), 397-406. <http://infoz.ffzg.hr/INFuture/papers/7-02%20Milosevic,%20Stancic,%20Usage%20of%20Virtualization%20Technologies%20in%20LTP.pdf>.

⁶ *Reference Model for an Open Archival Information System (OAIS)*. Blue Book (CCSDS 650.0-B-1) (Washington, DC: Consultative Committee for Space Data Systems, NASA, January 2002). <http://public.ccsds.org/publications/archive/650x0b1.pdf>.

act proactively, to enable replacement of components and concepts (that is why OAIS is a reference model, implementation is always based on selected technologies). Technological environments and solutions on the market should be seen as currently available tools and the archival experts' assignment would be to estimate what are the best tools at the moment that can be used in and for creators' recordkeeping programs. In this way, archival experts should assess the value and risks of archiving in the cloud environment and act according to their professional evaluation.

3. Transition to Cloud Services

In order to better understand the tendency of transition to cloud services one should consider their possible advantages over the earlier, i.e., still the most dominant, information infrastructure organization like server consolidation. Putting aside the advantages of availability of data from any place with an internet connection or availability through mobile or tablet clients, cloud services could offer (semi)automatization of digital preservation functionalities. For example, cloud services based on storage of creators' content could offer, for a start, some analytical functionality that can run through the content and provide information on formats and versions of the stored data in order to provide possible adequate reaction in time. If potentially obsolete file formats are found service provider could alert its clients that their content could become unusable. This analytical tool could enable the provider to offer additional service. This second step could be, for example, automated conversion of client's content into higher or more stable formats, migration to advanced media (when necessary), emulation (if necessary) and re-authentication or validation of authenticity of the content.⁷ That means that the cloud solution could (semi)auto regulate itself against obsolescence of its contents. These mechanisms could be tested and implemented and become parts of cloud services' best practices. Of course, the same approach could be implemented on the local storage, but having a service, in some cases outsourced, with fully implemented preservation mechanisms could be the reason to choose one cloud service provider over another.

It would be misleading to expect that creators insist on using just standardized formats for their content. On the contrary, they are usually creating, receiving and working with various and non-standardized (or not preservation-appropriate) types of content. Creators do not always possess specific knowledge of maintaining the content, and that is one of the reasons for paying external service. Entity of origin is not always able to ensure quality and sustainability of the content. Additionally, archival institutions that are monitoring some of creators, namely creators from the state, regional and public spheres, have little or no control over dislocated contents stored in the cloud services. If the creators delegate custody of their content to the providers and archival institutions have limited or indirect influence on preservation processes that take place in the cloud, it would be a step back from the already achieved level of custody. Archival community should recognize that expanding horizons of service providers is very important for preservation processes today. Archival community is required to be a key factor in preservation, to transfer methodology and to ensure maintenance of the preserved content. The

⁷ Hrvoje Stančić, Krešimir Pavlina, Arian Rajh, and Vito Strasberger, "Creation of OAIS-Compliant Archival Packages for Long-Term Preservation of Regulatory Metadata, Records and Dossiers," in *eTELEMED – The Fourth International Conference on eHealth, Telemedicine, and Social Medicine*, ed. Lisette van Gemert-Pijne, Hans C. Ossebaard, Åsa Smedberg, Sincalir Wynchank, and Piero Giacomelli (IARIA, 2012), 105-110; Hrvoje Stančić, Arian Rajh, and Krešimir Pavlina, "Long-term Preservation Solution for Complex Digital Objects Preserved as Archival Information Packages in the Domain of Pharmaceutical Records," in *eTELEMED – The Third International Conference on eHealth, Telemedicine, and Social Medicine*, ed. Lisette van Gemert-Pijnen, Hans C. Ossebaard, and Päivi Hämäläinen (IARIA, 2011), 13-21.

key actor in this case is not entity of origin itself, but the provider of preservation service. It would be easier to deploy adequate preservation services by influencing providers of services through their best practices and guidance documents then by influencing content creators themselves.

4. Research

Archival institutions shifted from custodial acquisition model, due to hyper production of (electronic) records, requirements which were necessary to meet to ensure the readability of files, and capacities of archival institutions' repositories, to post-custodial activities like supervision and consultation of categorized records creators and to commercial services like providing archival arrangement to creators. Categorized creators of public records, with the guidance of archival institutions, started to function as archives for their specialized archival holdings and data assets, but hyper production and infrastructure problem inevitably reached them too. Although storage becomes affordable, recordkeeping services and procedures within repositories become more specialized and more complicated for records creators. Bluin Jr. and Rosenberg described processes of enriching connotation of records from purely transactional (evidence of transactions and vested rights) to social and political, when they become accumulated in repositories of archival institutions.⁸ This is basically the goal of archiving. Preservation of records implies preservation of records and their organizational, business and wider social context. With prerogative of long-term preservation (LTP) of records, as long as they have to exist for legal and business reasons, according to records continuum model, and with difficulties in protecting and preserving authenticity of complex electronic objects, creators are more and more interested in outsourced archival services which can support protection and long-term preservation of their electronic archival holdings. Therefore, postcustody or delegation of custodial function has reached the next step.

Basic requirement that provider of archiving service in cloud environment have to meet is to ensure the confidence of internal and external users in entity of origin and its archival holding. That means that provided service is safe and stable. Provider of services should follow all professional and technical regulations and plans or at least enable long-term preservation practice over entrusted digital objects. Objects intended for long-term preservation should remain accessible by usage of methods such as migration, typed object conversion or other digital preservation methods systematized by Thibodeau,⁹ ideally performed within provided service. Provider of archiving service should be aware that the service is not reduced to providing mere repository for storing digital documents. Service should encompass proactive archival management of (complex) digital objects, their organizational context and provenance. It should guarantee continued usage of archived objects as authentic and trusted sources of evidence for creator and information for present and potential wider user community. Link between this basic requirement (of safe and stable environment that provides confidence in creator and its archival holding) and additional requirements is scalability and long-term preservation proactiveness of service. Most creators have legal or business need for preserving authentic digital objects as resources or results of their business processes, but no professional or technical knowledge how to perform these tasks in the long-term perspective. Although creators are rarely interested in renting just

⁸ Francis X. Blouin Jr. and William G. Rosenberg, *Processing the Past: Contesting Authority in History and the Archives* (Oxford, New York: Oxford University Press, 2011).

⁹ Kenneth Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," in *The State of Digital Preservation: An International Perspective* (Washington, D.C.: Council on Library and Information Resources, July 2002), 4-31. Accessed June 4 2012, <http://www.clir.org/pubs/reports/pub107/pub107.pdf>.

storage space, some creators have very strict requests related to care for authenticity and probative value of their information holdings and they require additional LTP mechanisms and services.

Providers of such services should offer response to creators' current needs, offer more than just dislocated storage, and become successful players in this new market. In order to provide adequate services they should satisfy requirements like high availability of cloud services, security mechanisms, protection of holdings and data assets like backup and recovery procedures, protection of records and data, contextual links with creators, protection of records and data authenticity, long-term preservation mechanisms, readability protection mechanisms, availability and restrictions mechanisms, distribution mechanisms, financing and insurance models that support long-term preservation etc.

5. Survey Results

The need for enumerated requirements has motivated us to conduct a survey in order to examine what is included in already available cloud services for (archival?) storage on the market today. The results that follow are showing the present state of affairs in private cloud infrastructure among the seven survey respondents: from a global corporation with offices in 58 countries (1), USA (1), Great Britain (1), Slovenia (1), Croatia (2 + 1 local branch of an international corporation).

As is to be expected, most answers fall into the "very familiar" category (see Figure 1) since all of the surveyed companies have "in production" implementations of private clouds as well as technical staff with the expertise needed to maintain and operate such systems.

The purpose of the private cloud is implicitly versatile, although it is apparent that the focus is either on supporting internal company operations or hosting external services rather than both (see Figure 2). The main reason for that lies in the fact that an implementation that would satisfy both types of users would create various access, security and maintenance issues.

It is obvious that hosting internal services, server consolidation and similar operational purposes are the mainstay of a private cloud infrastructure, however its extreme flexibility allows for adaptation to practically any purpose even when dealing with sensitive data. Data stored on the cloud include

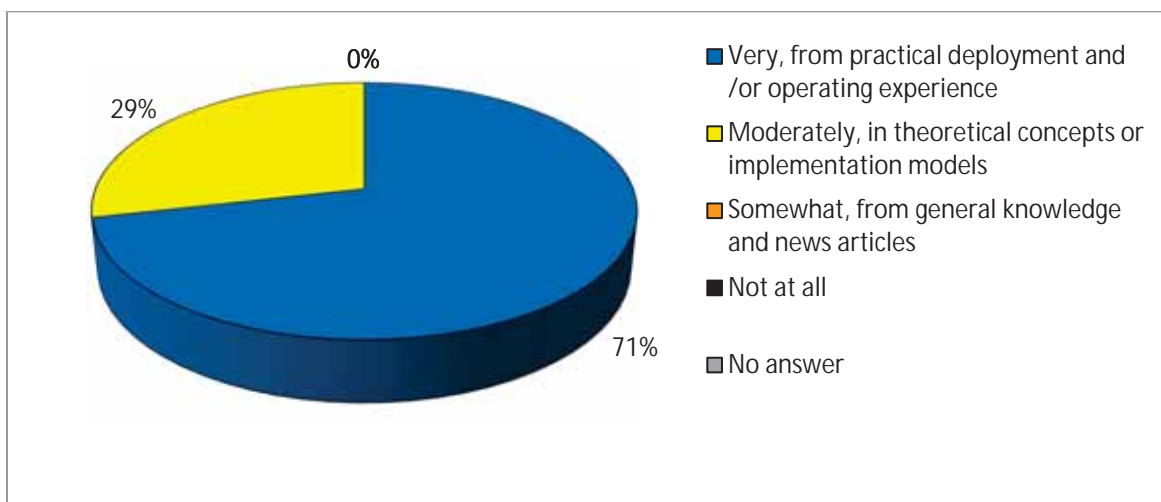


Figure 1. How familiar are you with the concept and/or technology implementation of private clouds?

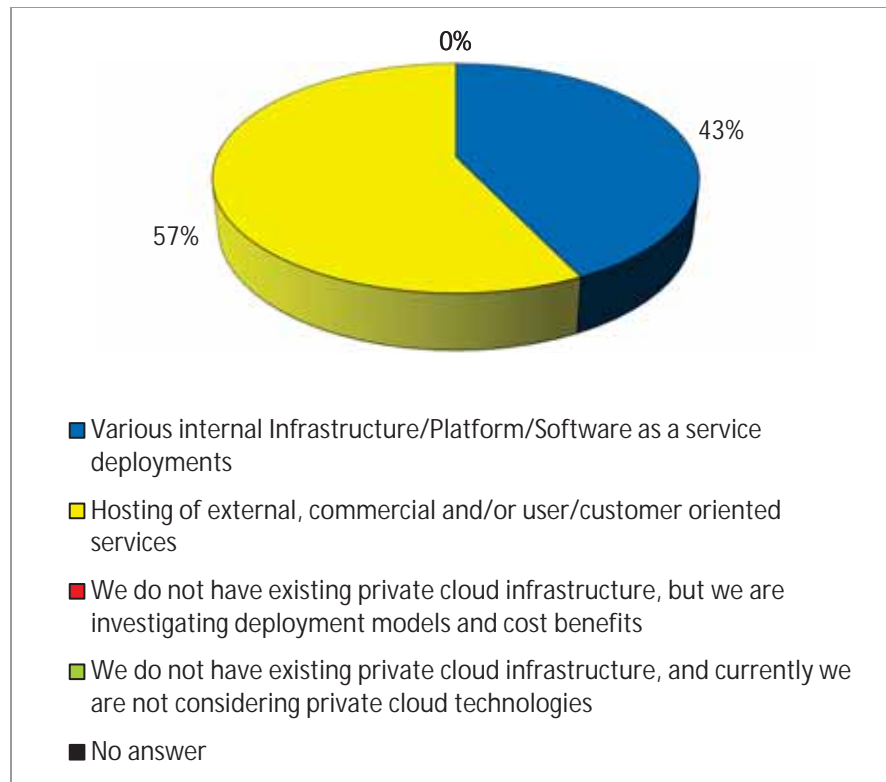


Figure 2. What is the primary purpose of your existing private cloud infrastructure?

information, for example, about employees with their personal data taken for HR purposes (see Figure 3). This shows confidence in security and authenticity mechanisms of the cloud infrastructure which is on par with more traditional IT infrastructure implementations.

In the “Other” category respondents indicated the usage of digital certificates and built-in access controls in a content management system. The comment: “We are not posting confidential information, such as contact information (...), online even with access restrictions” shows that the users are still careful in trusting cloud solutions. Nevertheless, it is obvious that customers and service providers are investing a lot of effort in data security in private cloud implementations since many of the proposed methods for data protection and access control are implemented in a majority of surveyed companies (see Figure 4).

Enterprise level cloud-based solutions for state or image backup are becoming increasingly reliable and have reached a maturity state where they are gaining ground to the more traditional OS-based backup solutions (see Figure 5). With their inclusion into standard enterprise license offerings these solutions provide rapid disaster recovery as well as file level and incremental backups, therefore becoming an all-encompassing recovery tool especially useful in private cloud implementations.

Some of the more advanced features of virtualization clusters and cloud infrastructures (such as HA, or DRS systems¹⁰) are obviously one of the main attractors to implementing services in the cloud

¹⁰ HA - High Availability - provides cost effective, automated restart within minutes for all applications in the event of hardware or operating system failures (<http://www.vmware.com/solutions/datacenter/business-continuity/high-availability.html>); DRS - Distributed Resource Scheduling - continuously monitors utilization across a resource pool

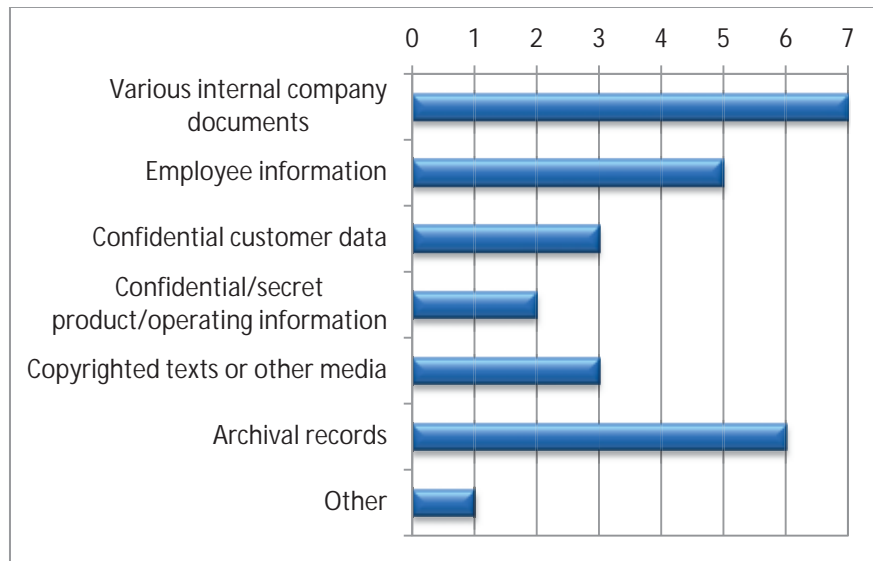


Figure 3. Does your private cloud implementation deal with sensitive data or copyrighted materials?

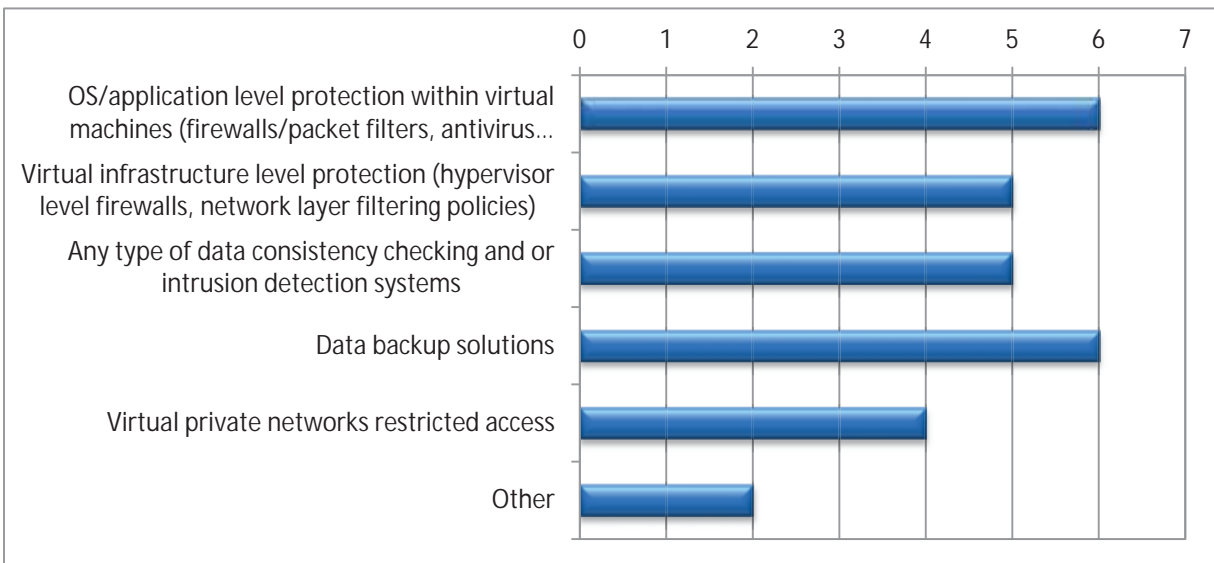


Figure 4. What methods for data protection and access control are you currently utilizing?

since many of them are readily available for installation with the selected virtualization software. It is not unusual that most of the interviewed companies have chosen to implement them in their private clouds. Some of the more costly ones requiring offsite or redundant datacenters are logically the “luxury” of financially more capable institutions while on the other end of the same scale, in the “Other” category, one respondent indicated usage of “backup stored on-site on labeled DVDs, with a README file on

and intelligently allocates available resources among virtual machines according to business needs (VMware vSphere, URL: <http://www.vmware.com/products/drs/overview.html>). Accessed August 8, 2012.

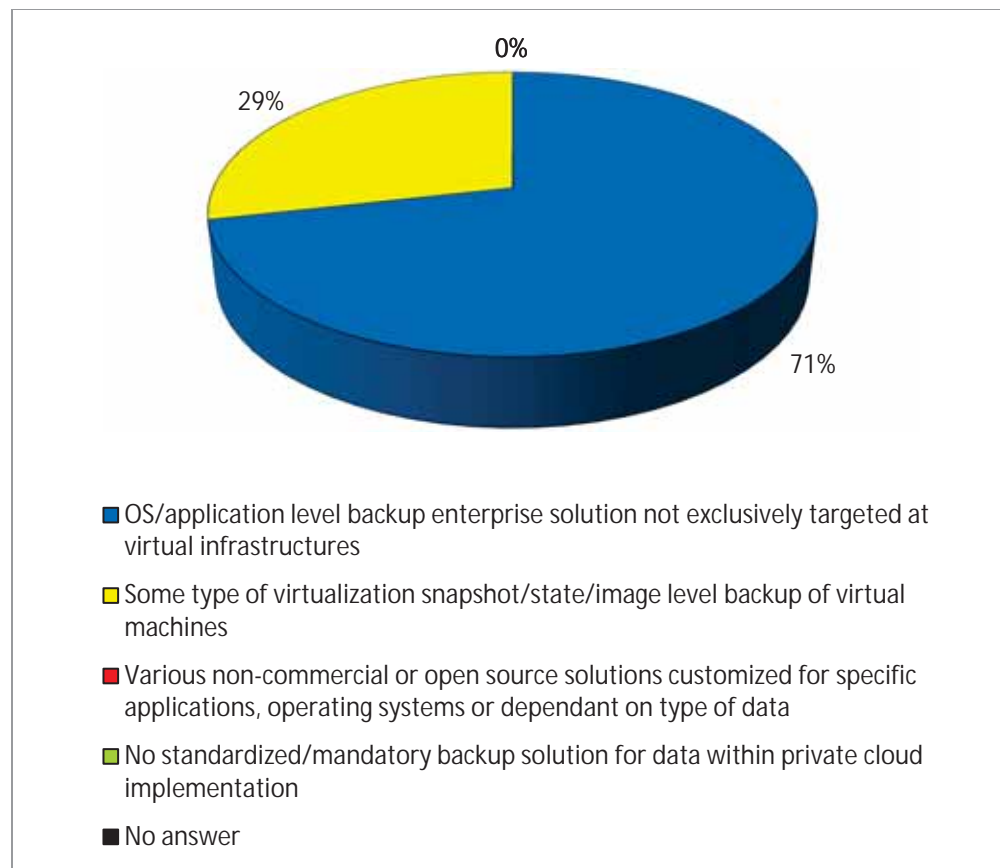


Figure 5. What type of backup technologies are you using within your private cloud?

installing the backup on a server.” The results show diversity of used approaches but still the trend towards the more advanced solutions can be clearly seen (see Figure 6).

The big issues with cloud implementations are security and usage policies. Although most replies state that there is some form of official policy in place, less than half represent a fully structured document with all the necessary specifications outlined in the appropriate manner (see Figure 7).

Recordkeeping implies written policies with implementation monitoring. ISO15489 recordkeeping standard puts more emphasis on policy making and documenting than on building technical part of the system itself. It also states that recordkeeping policies should be endorsed on managerial level and promulgated throughout the organization. That means that users, the key ones as well as the regular ones, should become familiar with recordkeeping policies and the answers from the survey show that the majority of employees are not familiar with them.

A great benefit of owning a private cloud is the implicit versatility. Respondents, for example, indicated their intention to implement an intranet in the cloud as well as a digital library (see Figure 8). Overall, the survey answers here show that although server consolidation is still the primary and most widespread way of information infrastructure organization, a private cloud is being utilized as well. However, depending on specific interests, all other areas of information infrastructure are included too.

Concerns regarding private clouds vary greatly according to specific institution needs and requirements (see Figure 9). Standardized or out-of-the box solutions rarely cover all concerns and com-

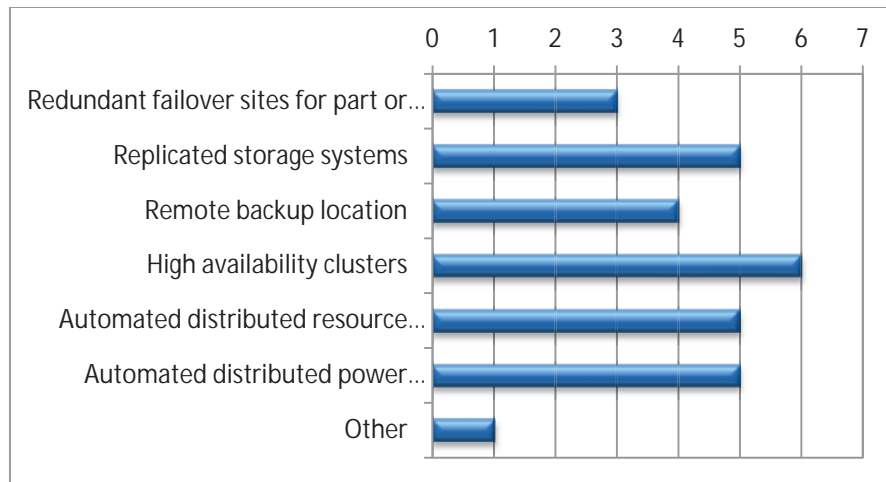


Figure 6. Do you currently have any of the following disaster recovery, accessibility technologies for your private cloud?

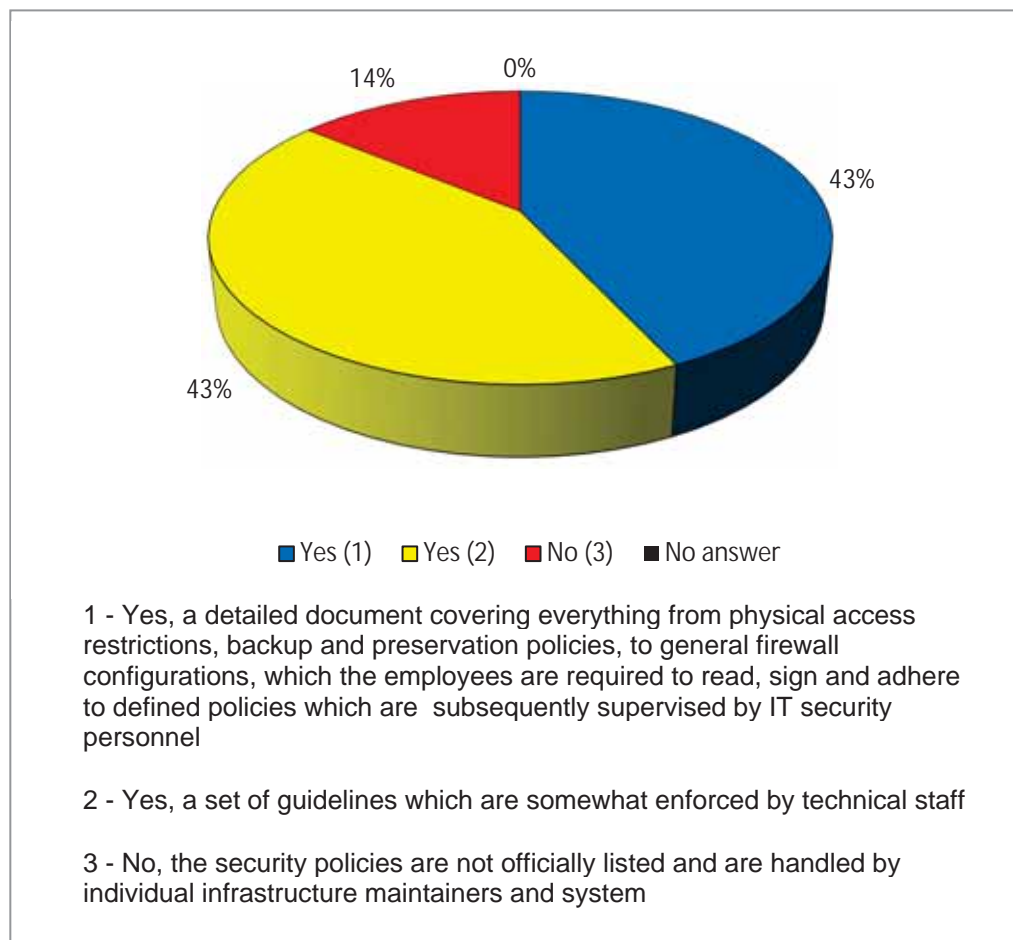


Figure 7. Do you currently have any type of security policy regulating infrastructure usage, access, password policies or long-term preservation of data?

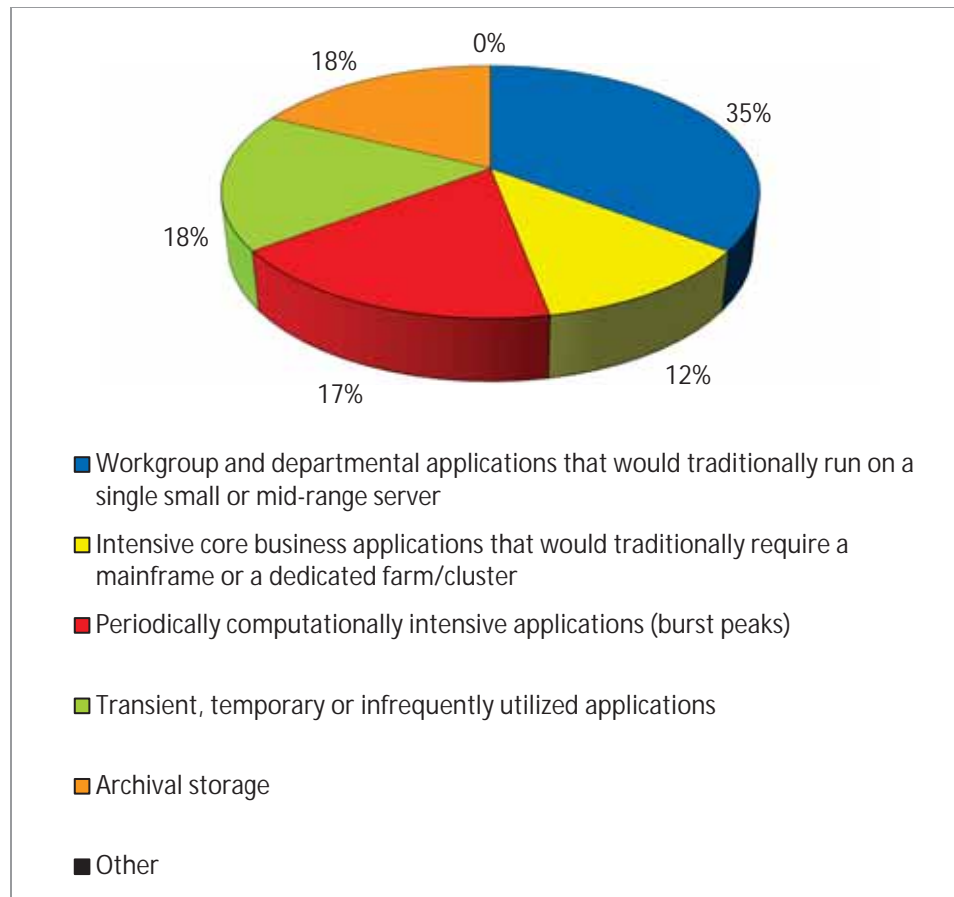


Figure 8. Do you plan to use or presently use your private cloud infrastructure for the following purposes?

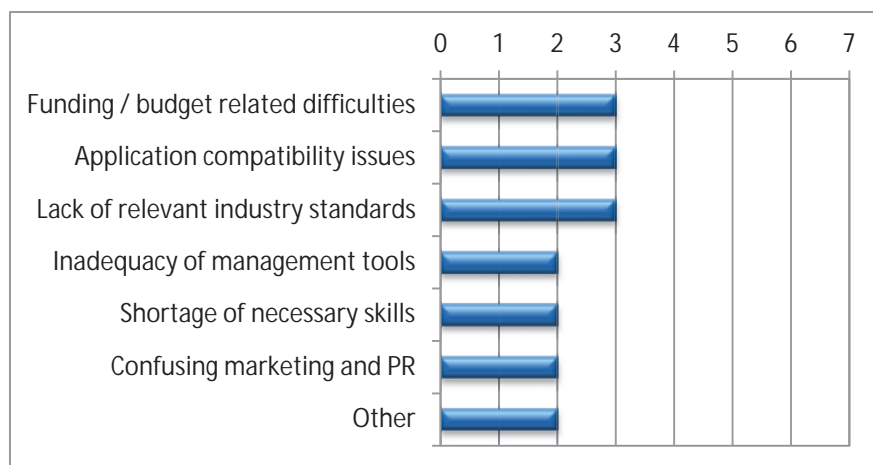


Figure 9. Would you agree with these frequently listed concerns about private cloud?

panies are often required to draw on internal technical staff in order to customize one or several solutions to their liking. Therefore private cloud implementers are still weary of the features, workload, skill and applicability problems with such technologies—analogous to their internal staff, operation procedures, technology capabilities and reliability. One respondent bluntly commented the concern for shortage of necessary skills: “There is not available talent to do something in-house which multiple staff can understand and maintain.” Lack of relevant industry and other standards is still a big issue and this gap could be filled by issuing best practices that could be followed by more and more providers in order to enhance cloud services and implementations.

The results show that cloud solutions are dominantly used for shared storage space and storage of specific type of content (see Figure 10). This means that the cloud is still not used as a full grown business solution but as expansion and/or upgrade of the needed services for certain types of content. From this it can be concluded that storage and archiving in cloud environments are upcoming services. This is issue that should be covered by archival professionals and their contributions could be included in best practice documents. Their skills in the fields of organization, indexing, classification and registration, as well as their preservation related knowledge and experiences, should be added and used in the design phases, i.e., before implementation of cloud environment itself.

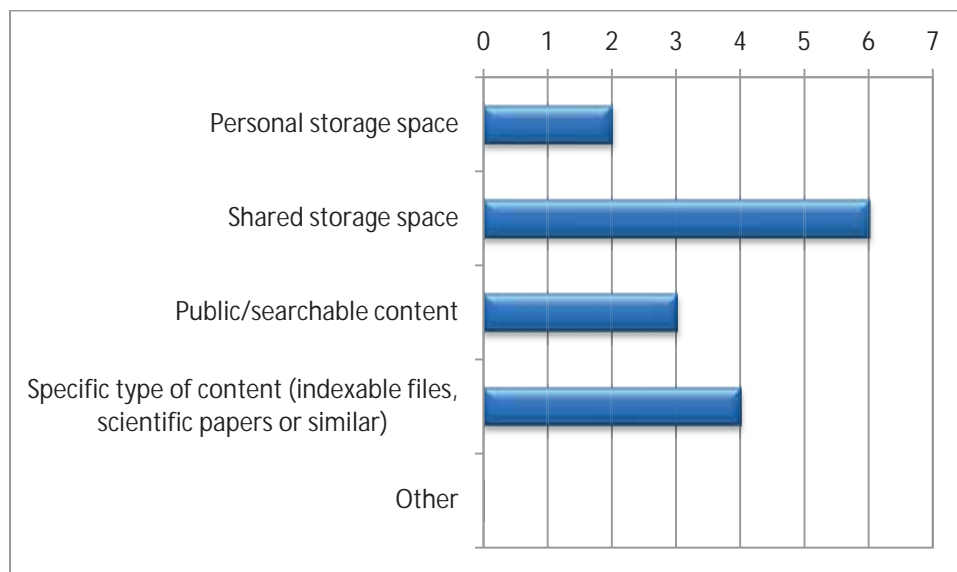


Figure 10. Does your private cloud host any type of data repository / archival storage service?

Interestingly, as shown in Figure 11, all surveyed institutions have one or more mechanisms for ensuring data authenticity or integrity implemented, and no institution is using any mechanism other than those listed in the survey. Under “Versioning” the usage of “audit trail with rollback” was indicated and “lots of provenance information” under “Metadata” category. This shows that users recognize the importance of content authentication and addition of metadata in order to provide long-term preservation and interpretability of content. The goal should be to preserve authentic content which could be used for evidence and wider-interpretable purposes. Carefully designed metadata is the key for long-term preservation. It is recommended that metadata are selected according to the relevant metadata standards and adjusted to the specific content. This is also the issue that should be addressed by the archival experts.

Both static and dynamic content were represented in the respondents' answers (see Figure 12). Static content, like documents usually created by word processors and office applications, is easier to preserve over the long-term while multimedia and dynamic interactive content require special methods of accumulation, organization and preservation.¹¹

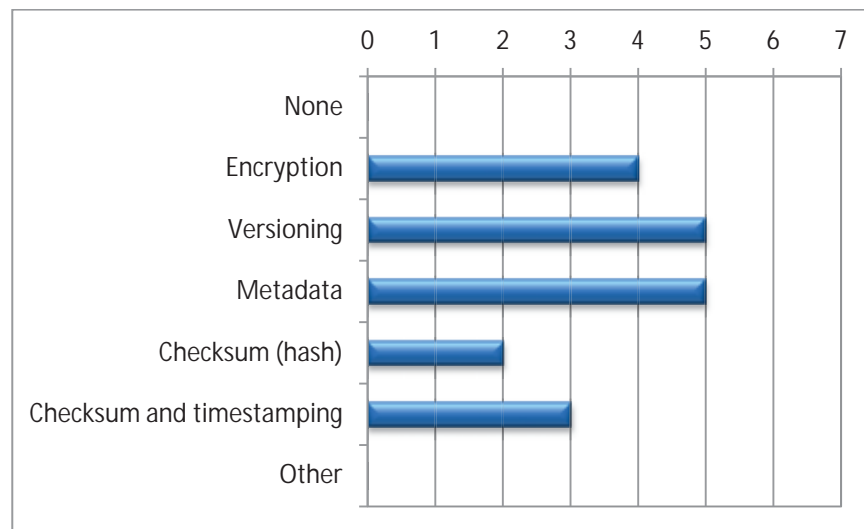


Figure 11. What types of mechanisms are used in order to ensure data authenticity/integrity?

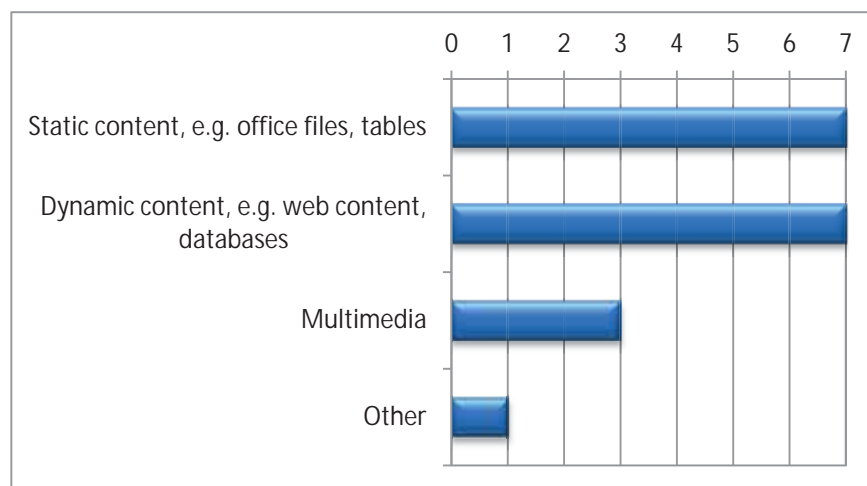


Figure 12. What type of content do you support with your service?

¹¹ For additional information on types of content see Martine Cardin, "Part Two—Records Creation and Maintenance: Domain 1 Task Force Report," [electronic version] in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records*, ed. Luciana Duranti and Randy Preston (Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008), 9. http://www.interpares.org/ip2/display_file.cfm?doc=ip2_book_part_2_domain1_task_force.pdf.

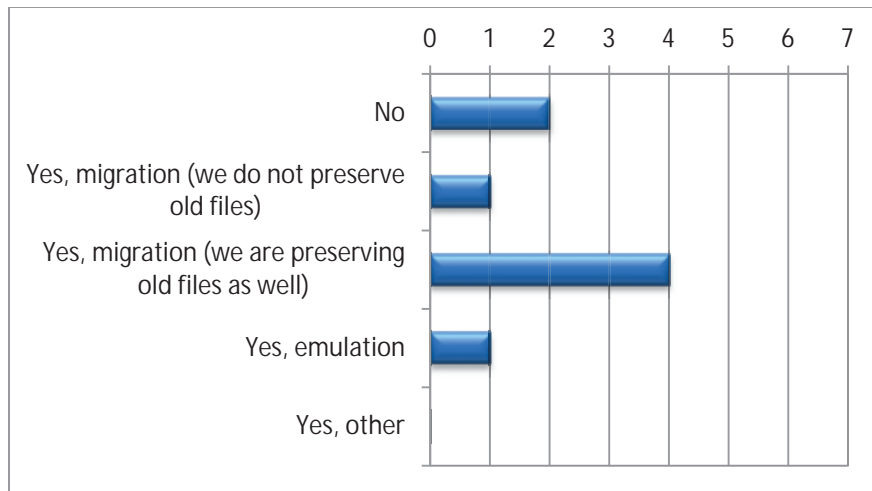


Figure 13. Do you offer one of the long-term preservation mechanisms as part of the service provided?

Although long-term preservation is indirectly very compatible with private cloud implicit technologies, it is in the process of being introduced as a cloud-based service. Some of the more explicitly applicable technologies are already being used for this purpose, while others are still being explored as alternatives or additional components of the cloud-based long-term preservation mechanisms. The results show that the importance of preservation planning is recognized among users and providers as well (see Figure 13).

Since cloud technologies imply indefinite scalability, all the surveyed companies recognized this and did not set or predefine a limit to the storage capacity which is to be allocated to their private cloud users (see Figure 14). This is how the cloud technologies should be implemented if financial and other considerations can be met.

As “Other” it was indicated that either the respondents do not charge for the service (3) or that there is a “fair use” limit for processing and a charge if exceeded (1) (Figure 15).

When asked to leave a comment on the survey one respondent indicated: “The two absolute best things for library repositories would be: 1) Awareness in the library community of the value of interoperable metadata, and 2) Easy to implement (i.e., built into content management systems) tools for using checksums to monitor for bit rot.” This comment from the users’ perspective excellently shows that the cloud service providers should cooperate more with the designated communities and experts in the field of archival science in order to be able to offer services tuned up to the specific archival needs and requirements, i.e., quality, preservation-aware services.

6. Discussion

The concept of archiving in the cloud environment or, as it could be called, “archiving-as-a-service (AaaS)” expands the definition of postcustodial archival practice in the new technological and organizational context. This could be seen as transition to the next stage of postcustodial archival practice. The notion of post-custody presumes that a creator is responsible for its own records and that archival institution supervises the creator. What could happen with the notions of responsibility and control during this technological shift? Who should perform supervisory activities, to what level do they extend and who

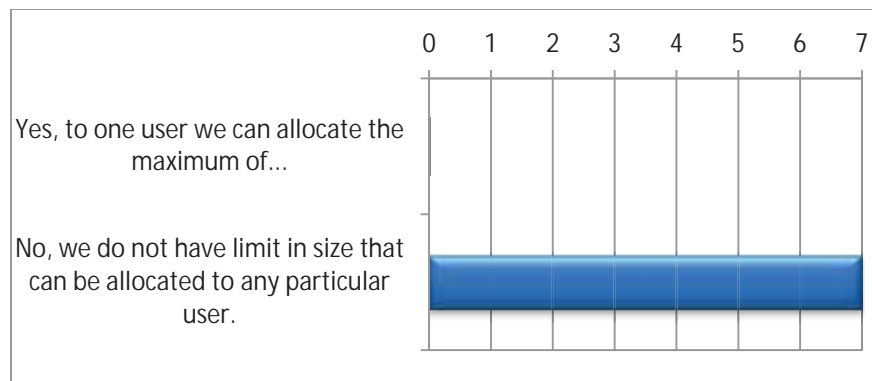


Figure 14. Is there any limit in the size (GB, TB ...) of the cloud storage that can be allocated to any particular user?

should be subjected to supervision? In the case of archiving in the cloud environment, where archiving practice is taken by service providers, several scenarios could occur:

1. Service providers are responsible for control of archived content without much interference and additional control taken by creators and archival institutions;
2. Creators invest much effort in additional control of non-standardized services;
3. Services are standardized through best practices and creators recognize the importance of choosing providers consistent to these practices;
3. Archival community is actively involved in the new concept of archiving and influence providers' practices.

The last scenario is probably the best way to ensure long-term protection, preservation and usage of electronic content created and archived today. This also implies updating of the concept of custody (see Figure 16). Characteristics of this new “postcustodial 2.0” stage would be comparable to positive shift in the public institutions' records management in the first half of 20th century when previously codified methodology entered into practice and made public records usable outside boundaries of their primary context.

7. Conclusions and Future Work

Custodial archivists from archival institutions had carried out custodial tasks on behalf of creators. Postcustodial archive monitored creators during creators' custodial activities. Today, custody is outsourced and archival community should direct its postcustodial advisory (if not supervisory) efforts to the new providers of custody. Therefore, it becomes much harder for the archival institutions to monitor and influence archival procedures. With the switch to the cloud services the records creating and preserving practice starts to elude the postcustodial influence of the archival institutions since it became technologically more difficult to achieve the sufficient level of supervision either over the quality of the services in the Archiving-as-a-Service (AaaS) approach or over the fully blown cloud service in which the records are both created and archived. Therefore, the archival community should persuade various recordkeeping service providers to prevent content obsolescence by adding additional functionalities such

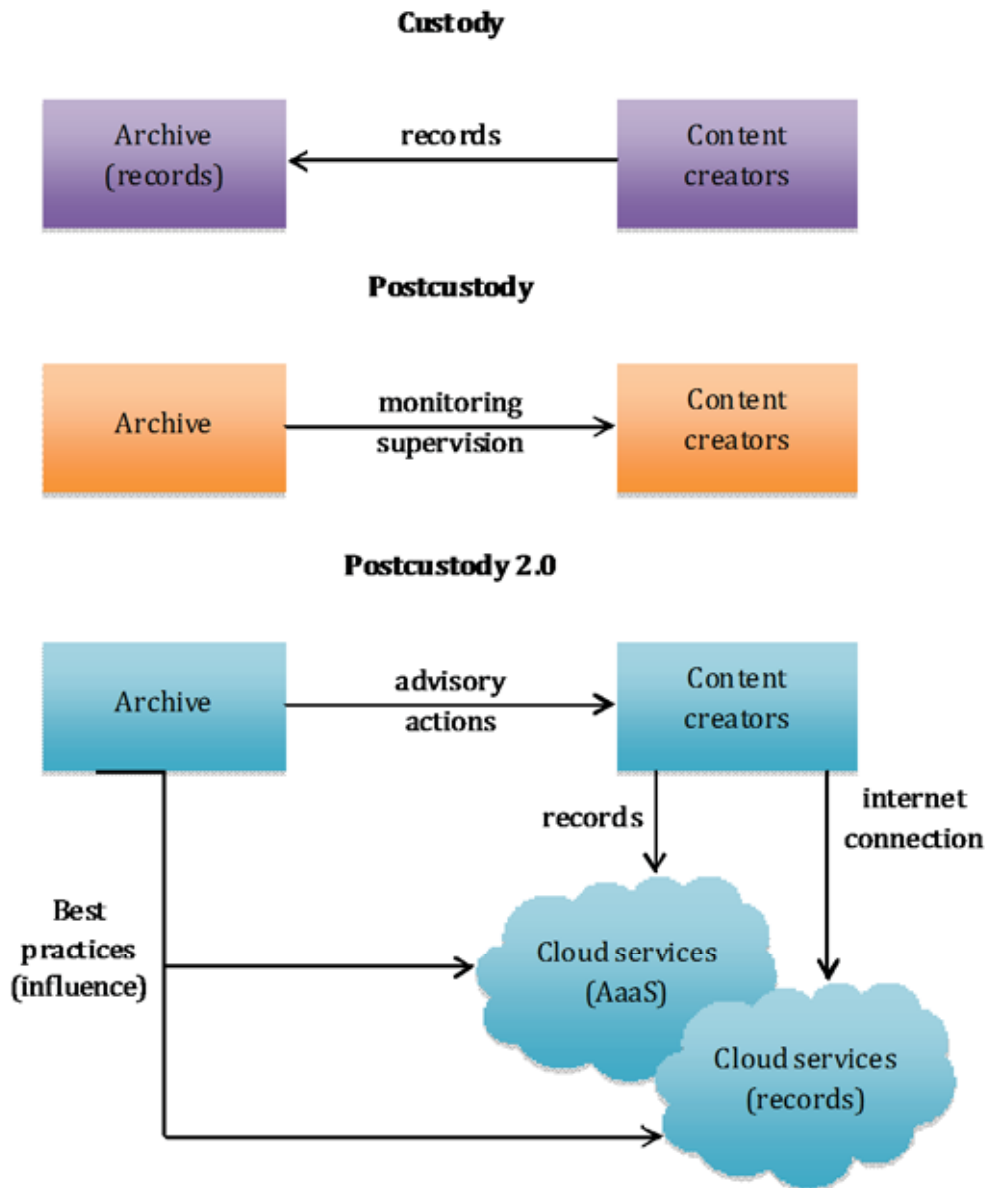


Figure 16. Changes of archival practice.

as content analyses and preparation for migrations and/or emulations. Wider approach could be an active involvement in production of best practice documents (and standards in time) for cloud implementations. Archival community should recognize the cloud computing discourse and contribute as well as add archival notions and terms to crucial zones of the new discourse. A good way to get archival notions implemented is through cooperation with experts in cloud computing and provision of cloud services, and through their best practices (in lack of standards for this particular area). The goal of archival community, and that is safe and secure long-term preservation of authentic, reliable and reusable content, should also be the target for providers of cloud services who want to rise above the competition and offer a better level of service. The archival community should therefore act proactively and impose itself as a key factor in the formation of the new, preservation-aware cloud services.

Planned future research and actions consist of detection of further trends in providing cloud services with implemented archival requirements as discussed in this paper. Also, the plan is to raise awareness both within archival community for the need to require archival level of cloud services as well as within cloud service providers for the need to implement the appropriate mechanisms. It is planned to include this problematic in the graduate level curriculum of the Study of archival science at the Faculty of Humanities and Social Sciences, University of Zagreb.

The CODATA Mission:
Preserving Scientific Data for the Future

Recovering the Forgettery of the World

Elizabeth Griffin and the CODATA DARTG Team

Dominion Astrophysical Observatory, Victoria, Canada

Abstract

This Session of the conference examines an aspect of scientific data from a somewhat unusual perspective, by focusing on observations that are not, or not obviously, a formal part of the memory of the World. Almost exclusively those data are historic records; they complement modern ones in a way that can be crucial for accurate studies of long-term trends. A CODATA Task Group, “Data-At-Risk”, is creating an Inventory of such data that catalogues their locations, types and volumes, and risk levels, as a prerequisite for designing efficient programmes to rescue the scientific information that they contain. The rationale for the CODATA effort is rooted in the knowledge that heritage data can offer unique evidence for solving some of the most pressing scientific unknowns that the world is currently facing. This Introduction examines that rationale, and the presentations that follow illustrate different types of data that are currently being rescued and brought into the public domain.

Author

Dr. Elizabeth Griffin is a volunteer researcher at the Dominion Astrophysical Observatory in Victoria (Canada). Since her Ph.D. at Cambridge (UK) in 1966 she has researched astrophysics at Cambridge, Oxford, Antwerp, Brussels, Toulouse, Boulder, Toronto and (now) Victoria. She Chairs an International Astronomical Union Task Force for preserving and digitizing astronomy’s heritage data, and also Chairs the CODATA DARTG mentioned here. One recent focus of her research has been the recovery and re-use of historic data for trans-disciplinary studies.

1. Preamble

The Memory of the World is a superb resource for scientific research that explores changes in the biosystems and the physical conditions of our planet. Many natural sciences are needing to call upon that memory nowadays in order to understand and quantify, in particular, the sorts of *long-term* changes that not only biosystems but also our physical world—its oceans, air, glaciers and deserts—are experiencing. Many of those changes appear to be substantially more radical than anything recorded during the last century, but the baseline observations which are needed for assessing the reality of trends are unfortunately not readily available for inclusion in research. Modern analyses require data to be electronic, so records that are not in that format are effectively inaccessible. That immediately excludes almost all historic data, because they were recorded on non-digital media such as paper, film, photographic plate, books or undocumented magnetic tape, and have never been transformed from that state. As a result, the very observations which are most needed to determine long-term trends cannot currently be incorporated into relevant research programmes. Modelling has become highly sophisticated, and increasingly efficient at representing the data that are supplied, but are not able to go beyond those limits with any confidence: extrapolation is unreliable in a chaotic world, and the only reliable solution is to capture in electronic form the information that is latent in so many of those pre-digital data. Though ambitious in many respects, as discussed below, such a programme will extend the publicly-accessible memory of the world backwards in time to encompass the decades when anthropogenic interference was far less vigorous than it is today.

But there are many challenges. Some historic data are deemed unusable, some are physically degrading, and all risk being destroyed through ignorance, even though the time-spans which they represent may be crucial for current research. This article introduces the topic and examines the general situation and outlook, while the following ones explore the matter from a number of different angles.

2. Two Hemispheres of Scientific Data

The power, capacity, speed and capability of modern computers has changed radically how today's researchers tackle large quantities of data and address major scientific problems. These developments are relatively new. A dozen or so years ago it was a tough challenge to import, let alone merge, data from disparate sources in order to produce a comprehensive model or extract a trend; today, current technologies are able to recognize patterns, correlations and correspondences between data that may have quite different fundamental properties.

These advances are timely indeed. The observations which are central to research into the natural world have become increasingly detailed as technology has evolved. Many modern data sets are large and complex, and their analyses involve interactive adjustments to complicated models that may require supercomputing power. The results that emerge can be used to predict how situations will change in the future. In principle, if the models are sufficiently reliable they can also be driven backwards to reproduce past conditions over a time-span that is significant compared to the recent changes that are irrefutable in an alarming number of situations. In practice, however, an extrapolation cannot have the degree of dependability that is needed, owing to the element of chaos in the natural world; the wide range of contributing factors and the subtle interplays between them can trigger the unexpected quite legitimately. A model may therefore only be valid for the time-span of the data from which it was generated.

Investigating the causes of long-term trends in the natural world has become urgent science. If recent changes are due to anthropogenic interference, the sooner each situation is understood and quantified the more likely it can be halted and even reversed. Determining the causes therefore needs observations from the decades when those anthropogenic interferences were much less widespread and vigorous than they are today. However, that immediately raises a problem, because few digital data-sets date back more than 25 years, and managed archives of most electronic data are even younger. The historical observations that are so badly needed are not in electronic form, so cannot be ingested into the modelling analyses without some specific transformation procedure. No matter what type of historic observations are called for, be it concentrations of stratospheric ozone, recordings of ocean temperatures, amounts and distribution of rainfall, counts and sizes of sunspots, bird migration patterns, frequencies of extreme weather events, behaviour of marine organisms, or a veritable host of many similar daily, seasonal or occasional events, if they were recorded on non-digital medium they will require transforming, and many may also need specialist knowledge to decode or interpret them correctly. It is a serious concern that the very data which contain vital clues to a correct understanding of how and why our natural world is changing are inaccessible to analysts, and will remain so without some major effort. They are poles apart from modern, all-digital data and managed data-bases. They are almost "lost" to the scientific world.

Fashion also plays a troublesome role. Fashion has dangerously close links with consensus, and what tends to win research funding is what captures majority interests or promises prestige for the team, laboratory or country. The resultant polarizations of data have been seriously unhealthy for the sciences that need to access historic data, and if some of those observations get discarded for whatever reason, the science loses unrepeatable measurements. Nevertheless, the pendulum of fashion does swing back. The

medical researcher, whose painstaking survey in the 1960s of patient cholesterol levels was scorned at the time, had fortunately stashed away the packs of Hollerith punched cards bearing the acquired measurements, and when they were recently re-discovered they were recognized as a gold mine of long-term data.

There is clearly significantly more to recovering heritage scientific observations than keeping artefacts from the past for their antiquarian interest. The physical form and format of the observations is of little direct relevance, except in understanding their quality and limitations. The medium is not the message; the scientific value lies primarily in the date-stamp associated with each measurement. Delving back into the past is a route to instant science, and while historic measurements may lack some of the detail of modern-day ones their uniqueness in time, and the possibility that they will supply actual records of the behaviour of a given property throughout an extended time-base, renders them of unrepeatable value.

While sterling attempts are being made in some sciences to make data rescue a major feature of the research effort, in others (and for all its advantageous infrastructure of sorted and catalogued heritage, astronomy is perhaps one of the worst culprits), the rescue efforts are left to the passionate amateur who feels deep concern at the possibility of losing historic data and who is prepared to volunteer time and effort to doing what is required. Though the efforts of the amateurs are absolutely essential, an unfortunate side-effect is to exacerbate the existing polarization between the new and the old.

One basic lesson to promote is therefore the *complementarity* of old and new: they may add information of differing quality, but interpretation relies upon both. Education will certainly help, but real-life examples of new science that was *only* possible with recovered data can be vivid teachers. What needs to be done?

3. The Dangers of Doing Nothing

There is no doubt that scientific evidence which is crucial for understanding the impact of anthropogenic interference upon planet Earth and its countless biosystems will be found in historic data. Lack of such evidence has allowed arguments that are based on opinion rather than fact, and the confusion that currently surrounds efforts to ameliorate certain situations owes a lot to news sensationalism. The longer we wait for the true facts to emerge, the harder it will be to halt the situation, or reverse it if that is what is required. But not all the problems are on a global scale, nor are all of the solutions enormously difficult to reach. There is the tragedy of the African farming family, newly settled in a region and intending to set aside extra food the following season to cope with the droughts which—they heard from somewhere—occurred every 10 years, but faced severe starvation because the actual period of the drought was 6 years, and was about to happen; the weather records were not suitably accessible. In a quite different setting, the environmentally-damaging effects of forestation with non-native species in the mountains that are sources of Cape Town's reservoirs only came to light when a small group of researchers digitized 73 years' worth of paper recordings of stream flow and uncovered a strongly positive correlation. Both examples have much to teach. In both situations the observations which could furnish essential knowledge existed, but were not accessible. The information that was needed was not complex and not overwhelmingly large to handle, nor was complicated machinery involved at any stage, just a focussed human effort. Doing nothing had inflated expense, and ruined lives.

Not infrequently, scientific measurements find application in more than one field, even in different disciplines, though it may not be the same analysts who recognise alternative uses, and the analyses need

not be simultaneous, or even close in time. Providing access to historic data in electronic form can yield unexpected side-benefits that are often far removed from the original purpose of the observations, so doing nothing denies opportunities for those initiatives. The astronomers who observed the spectra of hot stars in the near ultra-violet in the 1930s in order to study interstellar absorption features had no idea that their observations could be used to deduce the concentrations of the Earth's stratospheric ozone, and had they been able to bequeath fully digital spectra then, the data would have been seized upon at once by atmospheric scientists. Since digitization was not then an option, the spectra were left in their virgin state (but at least preserved), and it was three-quarters of a century before that ozone research could be carried out. Somewhat similarly, meteorologists are recognizing gems in the weather logs kept routinely by submarines or astronomical observatories, and atmospheric scientists see patterns in crop yields that tell as much about El Nino events as grain performance. We are now in a position to perform good-quality electronic transformations of all such records, but important background information will be harder to capture as time passes and less human memory can be tapped for the properties of the original data and their specificity. While it should not be assumed that *all* the historic information which might in principle be made accessible will contribute to a positive solution in some field, it is a sobering reflection that lives could be saved, deleterious situations avoided and damaging traits repaired by taking account of the huge wealth of heritage observational material that is presently in the world's "forgettery".

4. The Forgettery of the World

A forgettery is part and parcel of the human psyche, and a highly efficient archiving system for the brain to stash away images that may hurt, or surplus facts that are very infrequently wanted. It is also the convenient location for things unwanted; progress can only feel good if there is a forgettery to handle the casualties that got in the way. So with institutional science: technology and change are closely coupled, and new ideas should be seized while young even if the older technology and its output are not fully wound up. New technology engenders new expertise, but moving personnel from the older to the newer has the unfortunate repercussion that rather little expertise and resources remain for curating the data which were so central to yesterday's research. The world has an enormous "forgettery" of scientific records that it might once have kept for a purpose but has almost forgotten what that was. It has physical locations in attics, cupboards, closets, archive warehouses, private journals ... The challenge for today is to bring that portion of the world's forgettery into the world's growing memory, where it can offer irreplaceable contributions to today's scientific research.

Why is that still waiting to be tackled? A comprehensive programme to activate those forgotten pockets of the world's memory may seem dominated by practical difficulties, including dealing with unfamiliar data from unfamiliar equipment, and the necessary expertise may be diverse and difficult to locate, but all that is surely less challenging than (say) launching satellites to make observations of the sky or the earth, and certainly less costly. What it *will* require is a united determination to overcome the inertia which has trapped so much valuable material doing nothing in a forgettery, so the first step is to publicise and educate. That was the basic rationale for this Special Session at a UNESCO conference.

Recognizing the seriousness of the situation, CODATA (the *ICSU* Committee for Data in Science & Technology) has mandated a Data-At-Risk Task Group (DARTG) to seek out sources of non-electronic

data, and create an Inventory.¹ Time is not on our side; some of the materials to be recovered are already becoming unreadable—deteriorating magnetic tapes, recorded data without sufficient meta-data (information about the information), photographic records that have been deprived of the essential equipment to measure them correctly, or hand-written sheets whose ink is fading. CODATA is clearly late in arriving on the scene, yet the two essentials for a constructive data-rescue programme—a recognition of the need to study long-term trends (the “why”) and the capabilities of technology to manage the necessary tasks (the “how”)—have only recently moved close enough together to promise truly worthwhile returns for the efforts to be invested. The world’s forgettery can yet be rescued and put to very effective use.

5. Stages of Data Rescue

The requirements of a programme to rescue historic information are varied and demanding, but not insurmountable. Each calls upon a number of different expertises, beginning with the history of the relevant experimentation, equipment and associated personnel, and can ultimately involve a broad selection of different groups and skills, so each needs to be sympathetic to the overall objectives. The challenges of data recovery on a scale to return significant science is beset with fascinating problems that are rather rarely encountered in modern programmes. Discovering what is out there—somewhere, coping with the condition of what one finds, and arguing for the costs as part of a modern programme demand skills that involve communication, archival techniques and human interactions, and all need to be handled successfully in order to overcome the prejudice that is sometimes encountered in modern research, *viz.*, that anything that is old must be inferior. Deciphering notes in observation log-books is often highly domain-specific, and the best minds may no longer be around. Understanding the purpose of an experiment, and through it the limitations imposed upon the observations and thereby of their interpretation, will entail recourse to old publications and reports, few of which may be openly available (either printed or electronically). Converting the necessary elements of an observation into a fully-transferable electronic record demands a finesse that cannot be over-stressed, while the required meta-data (e.g., as in FITS headers) for seamless ingestion into a modern database are *sine qua non*.

Each step requires a clear vision, unambiguous procedures and well-tailored programmes with milestones, benchmarks and templates. In reality the data that one may get to work with will be in assorted condition, and each will need specialized treatment at some level, implying devoted resources. Each stage can produce unexpected challenges; it is hard to be fully prepared, and setbacks should always be allowed for in the planning time-line.

5.1 Tracking down “lost” data

The biggest hurdle is undoubtedly the initial one of communicating the incommunicable: getting researchers to discover what was put in storage probably long before their own arrival at the laboratory, observatory, library, archive or bureau in question. One approach is to follow up leads through observations that were known to have been made; those can result in unexpected discoveries too. Another is to ask people to go systematically through their storage areas and list materials that have been there for more than a specified period. Activating the trans-disciplinary potential of scientific data can sometimes trigger discoveries too; enquiries about weather records (for instance) may jog an astronomical

¹<http://ibiblio.org/data-at-risk/items>

observatory into reviving what *it* regarded as routine records for internal use only, but which may in fact be significant for environmental studies. At the other end of the scale are records generated by projects in Government laboratories and which may be protected for a period from actual destruction, but are regarded as space-wasters and are sub-optimally stored. Such caches may be the archetypal dishevelled heaps of paper, getting more scuffed when shifted and in constant peril of an order to “clear out” the cupboards or rooms when a certain age limit is reached. We may never know what potentially important records have already been lost in that way, but can be sure that it has happened and will happen again unless their scientific potential is understood. Moving a department into new facilities may have unfortunate side-effects if old records are regarded as mere trash for disposal, though it is often during such relocations that real finds are made.

5.2 The state of discovered data

Some heritage materials are formally maintained in tolerably good condition, and include most of the necessary meta-data (possibly as a catalogue), a description of the purpose and source of the data, and pointers to publications that may have ensued. “Records” may be original observations, in a variety of forms and formats, or may be the information from those observations transcribed (most frequently on paper or some type of magnetic tape) as “measurements”, either freestyle or in a pro-forma layout (e.g., a printed chart with blanks to be filled in regularly).

Most of astronomy’s worldwide collections of 3 million or so photographic plates, some dating back over 100 years, are in passable condition. But to make the data publicly available electronically requires specialized digitizing procedures entailing purpose-built equipment and trained personnel, and the resources needed for that are hard to find. In a somewhat parallel situation, the US National Climatic Data Center’s data modernization programme has a huge basement filled with files of hand-written weather records from worldwide sources. It will take a major effort to put the all information online, but the records themselves are in good condition. The Berlin botanical museum presents a different set of problems: a huge underground store houses *samples* of plant parts rather than observations of them; the archive is “live” inasmuch as new specimens are constantly being added, and the challenge is to capture electronically all the observations of the specimens, not only the meta-data but also the measurements have been carried out on them but never formally published. The very diversity of even these straightforward cases demands astute planning and design. Informing that planning is an immediate objective of DARTG’s Inventory of data at risk.

An example of well-stored but almost abandoned data is the collection of forms recording ozone measurements made in Oxford (UK) from 1933-57. Neatly bundled and ordered, the set had been almost untouched for half a century; manual transcription of the information and expert analysis filled in a major gap in recorded patterns of ozone changes during the last century.

In contrast, the IEDRO² video, *Historic Weather Data Rescue and Digitization*, displays untidy heaps of papers containing weather records, lacking even basic sorting or classification. Such heaps may get moved on, and with each move comes a loss of specific information regarding why they were saved, by whom, what they should contain, and where they originated. If the heaps (or boxfulls, closed containers or locked store-rooms) have been left undisturbed for a long time they might be found in a comparatively virgin state but deteriorating physically—ink fades, paper turns brittle, crumbles, or gets

²International Environmental Data Rescue Organization; www.iedro.org

attacked by mites or mildew, photographs discolour, emulsions become brittle and lift from their substrates, films crack, magnetic tapes oxidise and cannot be read—those are just some of the conditions that may await the investigator. A more unusual case of actual data loss is the sets of bolometric solar scans recorded on large glass plates from 1926-31 at a mountain site in Namibia; modern researchers could have mined them for information about variability in atmospheric constituents, but the plates were left abandoned on site and some have since found new life as window-glass in Namibian homes (after the offending wiggly lines were removed ...)

5.3 Influences of human attitudes

Natural ageing processes, even when speeded up by poor storage conditions and fungal infections and the like, are relatively slow compared to irreversible decisions by humans to jettison a collection because the space they occupy is wanted for some new activity, and it is sad to reflect how often the fate of such records is actually decided by ignorance. Collections have been destroyed “because no-one uses them,” but would they not be used if the information they contain were available electronically and the significance of the date-stamps properly appreciated?

The way scientific research is funded tends to exacerbate the problem. The concept of pushing a boundary backwards in time by digitizing heritage observations does not command the same prestige as building new equipment to carry out observations of world-breaking class, even though the latter projects are almost always the more expensive by a wide margin.

6. Pushing Forward

In order for its endeavour to succeed, DARTG needs to demonstrate and publicise the unique role which historic data throughout the natural sciences can very often play once they are made widely accessible in machine-readable formats. The Inventory which has been started will demonstrate the extent and types of *known* data that need to be the foci of rescue projects, whether for preserving, cataloguing or actual digitizing, and the more complete it is the better it can demonstrate the relative urgency and fragility of the various entries. DARTG’s ultimate objective is to extend the memory of the world backwards in time through a period that is scientifically significant, and conferences such as this can offer valuable platforms for broadcasting DARTG’s message and engaging broad participation by the community at large.

The excitement of scientific research is always enhanced by the discovery of the unexpected, and historic data have already provided results that were never anticipated at the time of their recording, either by those who made them or by those who now recover them. Keeping an open mind for opportunities of a trans-disciplinary nature is therefore key, both for maximizing the return on the recovery effort and for extracting new science that can stretch the imagination far beyond expectation. It is not ours to judge whether the wisdom thereby derived is more valuable because it teaches how to reforest hills, or keeps isolated farmers in touch, or measures the earth’s atmosphere through which we star-gaze, or clinches questions of climate change on a global scale. Every new turn in the zigzag path of scientific progress opens new vistas that are essential aids in an international quest to harmonize with our parent planet.

This contribution has provided some background for the Session on “data at risk”. The following three papers describe selected examples of work in progress in a variety of disciplines.

Tide Gauge Data Rescue

Patrick C. Caldwell

*Manager, Joint Archive for Sea Level, National Oceanographic Data Center (NODC), USA,
Patrick.caldwell@noaa.gov*

Abstract

Historical, non-digitized tide-gauge records are potentially of great value to the oceanographic research community as they can extend existing sea-level time series as far back as possible in order to understand more completely the time scales of sea-level change, and in particular, sea-level rise associated with climate change. At the 12th session of the Global Sea Level Observing System (GLOSS) Group of Experts (GLOSS GE XII, 7-11 November, 2011, United Nations Educational, Scientific and Cultural Organization, Paris), the topic of rescue of tide-gauge data in non-computer form (charts, tabulations, etc.) was discussed. The GLOSS GE acknowledged that a large amount of historical data remain in paper form and noted that there have been recent findings in non-oceanographic facilities such as the United States National Archives and Records Administration and the archives of the French territorial divisions. To learn more of the holdings of tide-gauge records worldwide, a questionnaire was developed and sent to national focal points for GLOSS and also to national hydrographic agencies identified via the International Hydrographic Organization. The questionnaire sought specific details on locations, time spans, sampling frequencies, and media type, volume, and quality. The responses were compiled in an inventory of the Committee for Data in Science and Technology Data-at-Risk Task Group, which seeks to assess the availability and quality of historical records from a wide spectrum of scientific and technological fields, with the long-term goal of identifying funding sources and means for transferring the old records into computer-ready format(s). This paper describes the accomplishments of the 2012 GLOSS questionnaire.

Author

Mr. Patrick Caldwell received a Bachelor's and Master's degree in meteorology from Florida State University in 1982 and 1984, respectively. He supported climate data rescue within the Marine and Environmental Protection Agency of Saudi Arabia from 1985-1986. He joined National Oceanic and Atmospheric Administration in 1987 as manager for the Joint Archive for Sea Level based at the University of Hawaii Sea Level Center.

1. Background

Objective measurements and observations are essential to the advancement of science. Within natural sciences, hypotheses on temporal and spatial variability of a given process are tested with objective theories and validated through *in situ* data. Various entities, such as national and international environmental data centers, support scientific research by maintaining large archives of readily-available, scientifically-valid, computer-ready measurements and observations. Long-lived entities, such as select museums, libraries, academic departments, non-governmental organizations, private environmental and engineering companies, and others within all branches of governmental agencies from local to federal levels, have collected and stored *in situ* environmental records. In some cases, data may be stored but are may risk loss owing to deterioration of media—or simply get forgotten. Determining what records exists can be just as daunting as the actual transformation into digital, science-ready form.

Since 2011, the International Council for Science: Committee on Data for Science and Technology (CODATA) has supported a Data-at-Risk Task Group (DARTG), whose purpose is to rescue scientific data. The international members of DARTG include data specialists in a wide range of natural and information sciences; each team member, upon whom DARTG may call, is an expert within a select discipline. The first phase is to gather information about data at risk. The collected information is placed into an on-line DARTG inventory (www.ibiblio.org/data-at-risk/items/browse/1). The motivation is to alert scientists to the existence of these valuable historical records. That awareness should facilitate funding solicitations by keenly interested researchers who seek to salvage safely the measurements and observations from their original media to contemporary electronic formats.

The author of this paper is a member of DARTG as well as the Group of Experts (GE) of the Global Sea Level Observing System (GLOSS). GLOSS was established in 1985 by the Intergovernmental Oceanographic Commission (IOC) of the United Nations Educational, Scientific and Cultural Organization (UNESCO). It provides oversight and coordination for regional and global sea-level networks in support of scientific research. The GLOSS GE has identified 290 tide-gauge sites worldwide; they constitute the GLOSS Core Network (GCN).¹ The site selection is based on several criteria; locations with minimal influence of rivers are desired to monitor better the oceanic variations. Preference is given to secure ports, which provide protection both from extreme waves and from vandalism). Sites with existing long time series are given priority.

GLOSS has designated data centers to support the securing of information about, and access to, tide-gauge measurements. Data centers are distinguished by the temporal turnaround from acquisition to on-line access and by the temporal resolution within the time series. Sea-level rescue has been an important focus of the delayed-mode centers: the British Oceanographic Data Centre (BODC), the Joint Archive for Sea Level (JASL) and the Permanent Service for Mean Sea Level (PSMSL). The BODC handles high-frequency series, defined as hourly intervals or less. The JASL, a partnership between the United States (US) National Oceanic and Atmospheric Administration's (NOAA) National Oceanographic Data Center (NODC) and the University of Hawaii Sea Level Center (UHSLC), focuses on data measured or reduced to hourly intervals, from which daily means are produced. The PSMSL is the longest-operating international sea-level repository, and has the largest number of sites and years for monthly mean sea level. All the centers acquire time series beyond the GCN. The centers share in solicitation from regional and national data suppliers, and exchange data and metadata on a regular basis.

Historical tide-gauge records are important for furthering our understanding of sea-level variations over a wide range of temporal and spatial scales. The highest frequency data provide guidance on magnitudes and durations of coastal inundations from extreme events such as tsunamis and storm surges. Regional, long tidal records have been used to study the temporal variation in tidal components.² Sea level has significant regional inter-annual through inter-decadal variations, such as seen in tide-gauge records from Hawaii.³ Historical data, through extension of the length of record for a given time series, augment statistical confidence of analysis. Studies of variations on those moderate time and space scales

¹ IOC, "Global Sea Level Observing System (GLOSS) Implementation Plan – 2012," *UNESCO/IOC Technical Series* 100 (2012): 2-48.

² Jay, David. A., "Evolution of tidal amplitudes in the eastern Pacific Ocean," *Geophysical Research Letters* 36 (2009): L04603, accessed August 1, 2012, doi:10.1029/2008GL036185.

³ Firing, Yvonne L., Mark A. Merrifield, Thomas. A. Schroeder, and Bo Qiu, "Interdecadal sea level Fluctuations at Hawaii," *Journal of Physical Oceanography* 34 (2004): 2514-2524.

and extension of records further back in time are essential for the task of defining long-term global sea-level rise.⁴

The Global Oceanographic Data Archaeology and Rescue (GODAR) project⁵ defines ‘data archaeology’ as the process of seeking out, restoring, evaluating, correcting, and interpreting historical data sets, and ‘data rescue’ as the effort to save data at risk by digitizing manuscript data, copying to electronic media, and archiving these data into an internationally available electronic database. In support of GODAR and GLOSS, several efforts have been made for sea-level data archaeology and rescue over the past two decades. During the 1990s the JASL salvaged 372 years for 34 stations of paper hourly tables, primarily acquired from the NOAA National Ocean Service (NOS) for sites within South and Central America.⁶ The BODC led a GLOSS archaeology and rescue project in 2001. Through this recovery effort and the work of individual agencies in digitizing and quality controlling paper records, 91 tide gauge series were extended backwards by 1,411 years of hourly data. The BODC has a substantial archive of approximately 3000 site years of tide-gauge charts and tabulations dating back to the 1850s. Some sites include other parameters. Most sites are within the United Kingdom and are in paper form. For eight of those stations, scanned images of the analogue tidal charts for 86 site years were made available, and of those, 45 years have been digitized. Funding has been secured to digitize 160 site years from 22 tide stations and produce scanned images of 500 site years for 14 stations, most of which date from 1890 to 1920.

At the GLOSS GE XII meeting in November 2011, the topic of data rescue was revisited. The primary motivation came from the inquiry of researchers Dr. David Jay and Dr. Stefan Talke of Portland State University (PSU), who have interests in historical records for analyses of tidal and other higher-frequency phenomena.⁷ Over the past several years they have inquired about the availability of historical sea-level records in need of rescue for tide-gauge sites within North America and the Pacific Ocean and under various international agencies. It has been determined that a large amount of records exist in non-digital form within the US National Archives and Records Administration (NARA) and the US Federal Records Center (FRC). In addition, Dr. Nicolas Pouvreau has recorded that large holdings also exist within the archives of the French territorial divisions.⁸ The GLOSS GE XII therefore decided to carry out a new inventory exercise about international holdings of sea-level data at risk, by issuing of a questionnaire to all GLOSS focal points and member representatives of the International Hydrographic Organization (IHO).

⁴ Church, John. A. and Neil. J. White, “A 20th century acceleration in global sea-level rise,” *Geophysical Research Letters* 33(2006): L01602, accessed August 1, 2012, doi:10.1029/2005GL024826.

⁵ Levitus, Sydney, “The UNESCO-IOC-IODE “Global Oceanographic Data Archaeology and Rescue” (GODAR) Project and “World Ocean Database” Projects,” *Data Science Journal* 11 (2012): 1-26.

⁶ Caldwell, Patrick, “NOAA Support for Global Sea Level Data Rescue,” *NOAA Earth System Monitor* 14-3 (2003): 1-8.

⁷ Talke, Stephen, David A. Jay, Patrick Caldwell, and Mark Merrifield, “Historical Tide Measurements in North America and the Pacific,” poster (2011), Civil and Environmental Engineering Department, Portland State University, Portland, Oregon, USA.

⁸ Pouvreau, Nicolas. “Three Centuries of Tide Gauge Measurements in France: Tools, Methods and Tendencies of Components of Sea Level in the Port of Brest” (Ph.D. diss., University of Rochelle, France, 2008).

2. The Questionnaire and Responses

The questionnaire was crafted through guidance of GLOSS GE and DARTG. It was sent out in early January 2012 with a deadline for May 1, 2012 (later extended to August 1, 2012). The purpose was to build an inventory of information about sea-level data in need of rescue. The primary desired information were where records reside, for what stations and dates, on which type of media, and in what condition in terms of readability and risk of loss. Specifics were also sought about the types, makes and models of the tide gauges, the recording mechanisms, clocks, data reduction, calibration, geodetic leveling, measurement of ancillary environmental parameters at the tide station, and the availability of technical, maintenance and processing notes. It is important to learn if the historical benchmarks can be linked to the existing geodetic network for a given station. A rough estimate of the volume of the physical storage media was requested. Additional questions concerned the original purpose of collecting the data, and whether copies reside in other repositories. Inquiries were also made about possible plans by the data holders to digitize the records in the near future, and if not, whether there would be scope for collaboration with other agencies/institutions to inventory and possibly rescue the data.

There was a total of 18 replies from 14 countries. Not all replies resulted in the discovery of historical data, though several mentioned that further investigations are ongoing. From the responses, nine repositories were identified as holding historical records: the Canadian Hydrographic Service (CHS), Danmarks Meteorologiske Institut (DMI), FRC, Instituto Geografico Nacional de Espana (IGN), Land Information New Zealand (LINZ), NARA, Rijkswaterstaat Waterdienst Netherlands (RWS), Servicio de Hidrografia Naval Argentina (SHN), and United Kingdom Hydrographic Office (UKHO). There is a total of 169 tide gauge stations (Figure 1, Appendix A) holding hourly or higher-frequency data at risk, and of those, 23 are within the GCN (Table 1 and Figure 2). The extensive historical sea-level data holdings identified in French repositories by Dr. Pouvreau are not included in this summary, since they are already well documented. The Tbilisi State University (TSU) also reported data holdings from three sites in the Black Sea, with a total of 126 years of monthly mean sea level. (High-frequency data are generally preferred as they allow for a greater degree of quality control and a wider range of applications).

The largest concentrations of stations are in Europe, North America and New Zealand, with a sprinkling of sites in Africa, Asia, Pacific Islands and the Caribbean. Only one site was identified in South America. The total time-span of those records adds up to 4,103 years, though excluding known gaps reduces the total to 3,259 years. (It is likely that there are additional gaps, so that number is still biased high). Data from some of those years have already been rescued and reside in GLOSS data centers. If digitized and quality controlled, these historical records could add 2,824 years of hourly data to the JASL, and 1,897 years of monthly mean sea-level data to the PSMSL. For GCN sites, they would add 324 years to the JASL and 270 years to the PSMSL.

The available non-digital records are of varying time spans (Table 2 and Figures 3 and 4). The time spans for the seven Danish sites are more than 80 years, though the degree of missing spans has not been determined. The FRC holdings for lengths greater than 30 years are most likely to have major gaps. The number of sites with newly identified series lengths greater than 30 years is 40, which represents 24% of the total, and of those, 35 series could be added to the JASL and 24 to the PSMSL. For GCN sites, rescuing those records could add spans longer than 30 years for 5 series to the JASL and 3 to the PSMSL.

The questionnaire included a number of items that could provide general technical information about the historical records (Table 3). The majority of the available historical records are in the form of analogue traces. Such pen traces are also referred to as marigrams, tide graphs or tidal charts. Only about

30% of the records have already been tabulated to numerical form as hourly or high/low-tide values. The vast majority of the media used is paper; only one station used film. The media and readability are good for 40% of the sites and of varying quality for 52%, while only 2 sites were described as poor. There were no confirmations that hard copies of the records were stored in other locations. The tide-gauge type was the standard float and stilling well for 42% ; only 4% used pressure or siphon gauges. The rest had an unconfirmed gauge type, though it was assumed most were of the float/well type. Station maintenance notes were documented for 45% of the sites; only one site was without any, but the rest were unconfirmed. Ancillary parameters of temperature, atmospheric pressure, and wind were taken at three sites; 15 sets did not having other measurements, and the rest gave no information. The inquiry as to why the data were originally collected gave the main motivation as being hydrography for port operation, including tide predictions and determinations of mean sea level to define data for navigational charts. Others noted geodesy in general, which could apply to near shore or land-based applications such as defining regional or national data. For the Danish sites, the need for knowledge about extreme water levels was also noted as a motivation.

Linking the gauge data to a vertical reference level is essential for most scientific applications. Daily visual tide pole or staff readings were historically the primary means of calibration. The tide staffs are linked to a network of land-based benchmarks through periodic geodetic surveys to determine any vertical movement of the station platform and to link the tide data to regional or national geodetic data, which could be tied to the same benchmarks. Tide-pole readings were confirmed for 46% of the sites, and only one site was noted as having only some sets available; no sites were declared void of readings. For the unconfirmed set, it is assumed that most have such readings since it has always been standard practice. Benchmark maps are available for 31% of the sites, though the rest were unconfirmed except for one. Historical geodetic surveys taken at the time of data collection were reported as being available for 40% of the stations, with only one confirming that none was available; the rest were unconfirmed. For 45% of the sites, the historical data can be tied into the present geodetic network, though most were unconfirmed. All of the agencies confirmed their support to GLOSS for access to these historical records.

3. Concluding Remarks

The GLOSS 2012 data archaeology and rescue questionnaire determined that a vast amount of historical tide gauge measurements exist in non-electronic form. Those measurements augment the large summary identified in French repositories and documented by Dr. Pouvreau. However, it is also recognized that participation in the questionnaire was only moderate, considering the large number of national contacts representing GLOSS and IHO. Among the replies, several contacts mentioned that an investigation is pending. Part of the reason is that the records for a given nation probably reside in disparate locations which are not readily near the GLOSS or IHO contacts. Thus, one conclusion from this effort is that additional searches of repositories need to be undertaken; that would require support or funding from national or international entities, plus willing national scientific/hydrographic/historical champions to perform the task.

One possible avenue for enlarging the effort could be to coordinate with other international groups that have similar goals, such as the Atmospheric Circulation Reconstructions over the Earth (ACRE) program. It is likely that many of the repositories holding data of interest to ACRE could also have sea level records. Thus, for example, if an individual searching for atmospheric data came across sea-level data, then a note could be made, and vice-versa. Other lessons learned in data archaeology and upcoming

plans of ACRE could be shared with GLOSS through collaborations. It is expected that a representative of the GLOSS GE will participate in the ACRE workshop in November 2012 (Toulouse, France).

The GLOSS questionnaire revealed that 24% of these discovered records are for time series with lengths greater than 30 years. Recovery of those data into scientifically-valid forms would add substantial lengths of record for many series. Such data could enhance the confidence in assessment of long-term global sea-level rise. Many other applications are possible for shorter times and/or more regional space scales, such as case studies of extreme events. The information from the GLOSS questionnaire will be made public through the GLOSS communication channels and the DARTG inventory. The inventory will be updated as new discoveries are made or if more detailed information is made available regarding missing years, such as in the case of the unknown gaps in the Danish and FRC sites. This study represents only the first stage of discovery of potential sources. It is hoped that this information will fuel the interest of researchers willing to seek funding and support for salvaging these records into computer-ready, high-quality data.

Acknowledgments

Thanks are due to Philip Woodworth, Lesley Rickards, Guy Woppelmann, Thorkild Aarup, David Jay, Stefan Talke, Sydney Levitus and Elizabeth Griffin for helpful comments during the preparation of the questionnaire and for constructive reviews of a draft of this paper. Stephen Shipman (IHO) is recognized for his support in sending the questionnaire to IHO national hydrographic contact points. Thanks are also due to the individuals and agencies that replied and provided information to the survey: Stefan Talke and David Jay (PSU, US), Shigalla Mahongo (Tanzania Fisheries Research Institute), Koos Doekes (RWI, Netherlands), Maria Jesus Garcia Fernandez (Spanish Oceanographic Institute), Puyol Montserrat Bernat (IGN, Spain), Afranio Mesquito (University of Sao Paulo, Brazil), Cesar Borba (Naval Center for Hydrography, Brazil), Juan Fierro (Hydrographic and Oceanographic Service of the Chilean Navy), Giorgi Metreveli (TSU, Georgia), Christoph Blasi (German Federal Institute of Hydrography), Palle Bo Nielsen (DMI, Denmark), June Thompson (UKHO, United Kingdom), Anne Ballantyne (CHS, Canada), David Wyatt (IHO, Monaco), Glen Rowe (LINZ, New Zealand), Angora Aman (University of Cocody, Cote d'Ivoire). Finally, thanks are given to NOAA and UHSLC for providing daily support for the author.

Tables

Table 1. A summary of GCN sites for recently identified non-digital sea level records.

Source	GLOSS ID	Station Name	Country	Time Span	# Yrs	Exist JASL 2012	Yrs to JASL
UKHO	0246	Cascais	Portugal	1905	1	1959-2005	1
UKHO	0248	GIBRALTAR	Gibraltar	1961-99	38	1961-2000	0
UKHO	0066	Honiara	Solomon Islands	1957-1961, 1965-1968	9	1974-2009	9
UKHO	0259	Lagos Bar	Nigeria	1940-1949, 51-53,69-70	15	1961-70,90-96	13
UKHO	0229	Reykjavik	Iceland	1956-63(part)	8	1984-1999	8
UKHO	0258	Tema	Ghana	1963 – 1964	2		2
IGN	0243	A Coruna	Spain	1950-1983	34	1943-2008	0
FRC	0290	Newport, RI	USA	1844-46,92-95	7	1930-2011	7
FRC	0220	Atlantic City, NJ	USA	1911-1939	29	1911-2011	0
FRC	0216	Key West, FL	USA	1847,50-52, 57-59,1903	8	1913-2011	8
FRC	0289	Fort Pulaski,GA	USA	1851-52,89-92	6	1935-2011	6
FRC	0288	Pensacola	USA	1890-1939	50	1923-2011	34
FRC	0217	Galveston, TX	USA	1852-1939	88	1904-2011	52
FRC	0159	La Jolla, CA	USA	1924-1939	16	1924-2011	0
FRC	0158	San Francisco, CA	USA	1853	1	1897-2011	1
NARA	0154	Sitka, AK	USA	1893-97, 1924-25	7	1938-2011	7
NARA	0206	San Juan, PR	USA	1892-1897,99	7	1977-2011	7
NARA	0073	Manila	Philippines	1901-1940	40	1984-2008	40
NARA	0116	Truk, Caroline Is.	Fed. St. Micronesia	1948-1949	2	1963-1991	2
NARA	0108	Honolulu, HI	USA	1877-1884, 1892-1905	20	1877-1892, 1905-2011	7
LINZ	0101	Wellington	New Zealand	1887-1944	58	1944-2010	57
LINZ	0127	Auckland	New Zealand	1899-1902	4	1984-1988	4
CHS	0156	Tofino, BC	Canada	1905-1908, 1917-1948	35	1963-2010	35
CHS	0155	Prince Rupert, BC	Canada	1906-08, 1919-1942	26	1910-1918; 1963-2010	26

Table 2. (A) A summary of the site counts (#) and percent of total (%) as a function of series length. The length excludes gaps. “All” pertains to the cumulative discovery of records. “JASL” refers to sites and years that potentially could be added to the JASL and similar for “PSMSL”. (B) This table follows the same theme though exclusively for GCN sites.

(A)		<=5 yr		5>yr<=15		15>yr<=30		30>yr<=60		>60 yr	
	sites	#	%	#	%	#	%	#	%	#	%
All	169	52	31	52	31	25	15	27	16	13	8
JASL	159	54	34	48	30	22	14	24	16	11	7
PSMSL	134	51	38	46	34	13	10	20	16	4	3
(B)		<=5 yr		5>yr<=15		15>yr<=30		30>yr<=60		>60 yr	
	sites	#	%	#	%	#	%	#	%	#	%
All	24	5	21	8	33	4	17	6	25	1	4
JASL	20	5	25	8	45	1	5	5	25	0	0
PSMSL	15	4	27	7	47	1	7	3	20	0	0

Table 3. Summary of questionnaire responses, as site counts.

Form of Data	Analogue Trace	Tabulated			
	118	51			
Storage Media	Paper	Film			
	168	1			
Media Quality	Good	Varies	Poor	Unconfirmed	
	67	88	2	18	
Stored Elsewhere	No	Unconfirmed			
	80	89			
Gauge Type	Float/Well	Pressure/Siphon	Unconfirmed		
	71	7	91		
Tide Pole Readings	Yes	Some	Unconfirmed		
	77	1	91		
	Yes	No	Unconfirmed		
Maintenance Notes	76	1	92		
Ancillary Data	3	15	151		
BM Maps	53	1	115		
Historic Geodetic Surveys	67	1	101		
Link to Present BM	76	1	92		
Cooperate GLOSS	169	0	0		

Figures

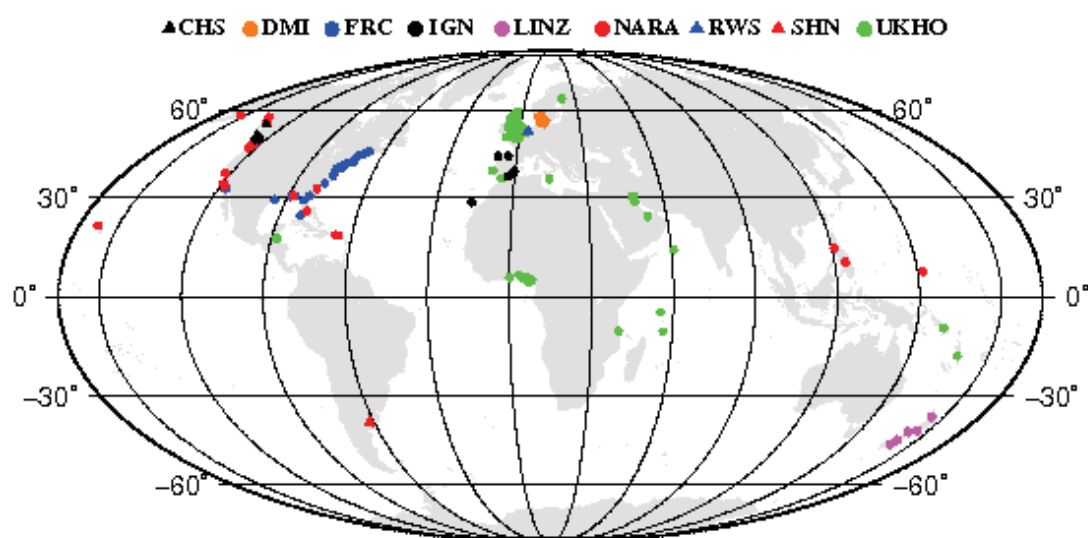


Figure 1. Plot of the station locations by repository for identified historical data in need of rescue.

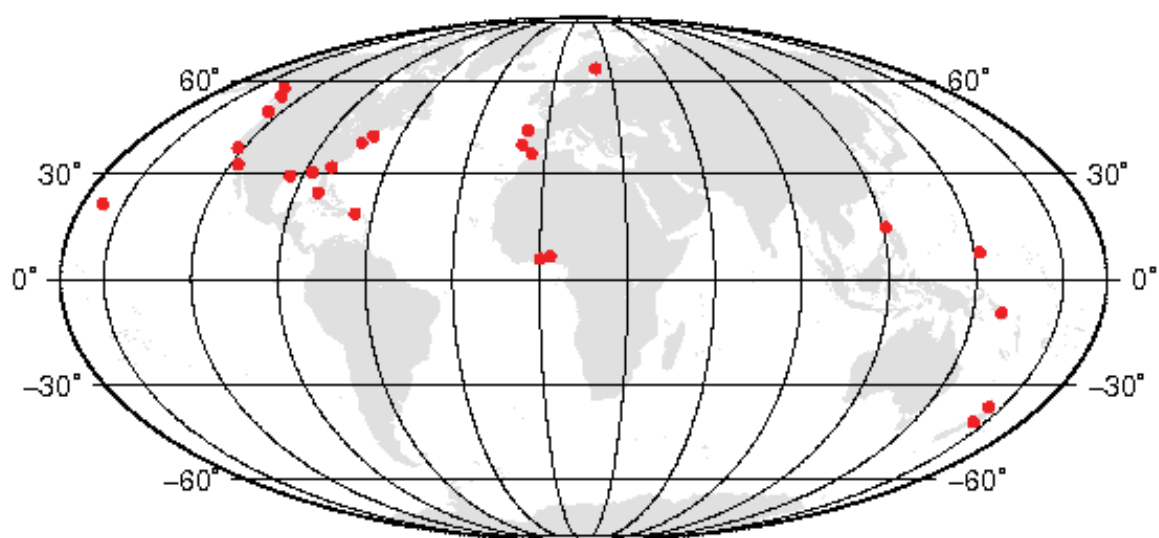


Figure 2. Map showing the locations of GCN sites with identified historical sea level data.

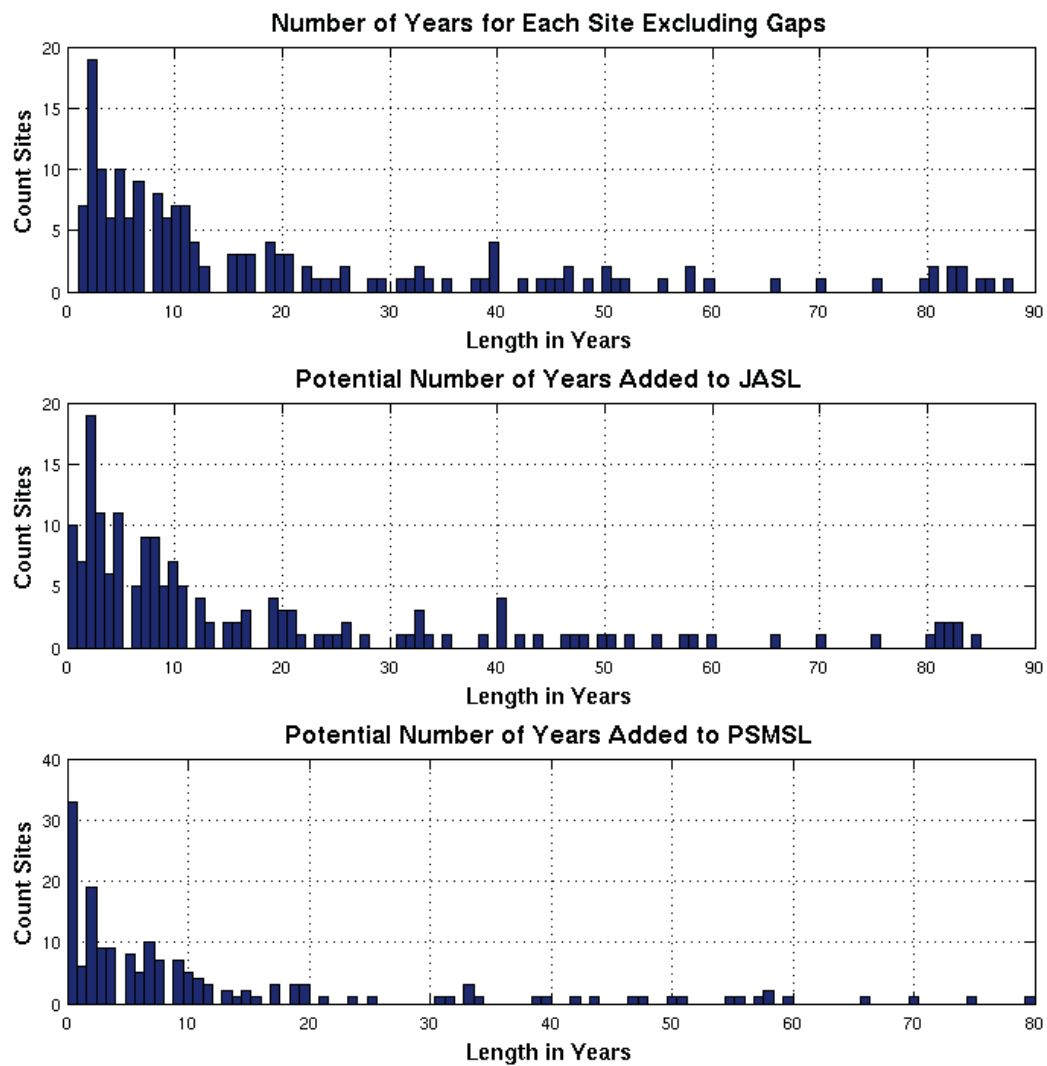


Figure 3. The site counts as a function of record length for the entire set excluding gaps, for the potential additions to the JASL, and same for PSMSL.

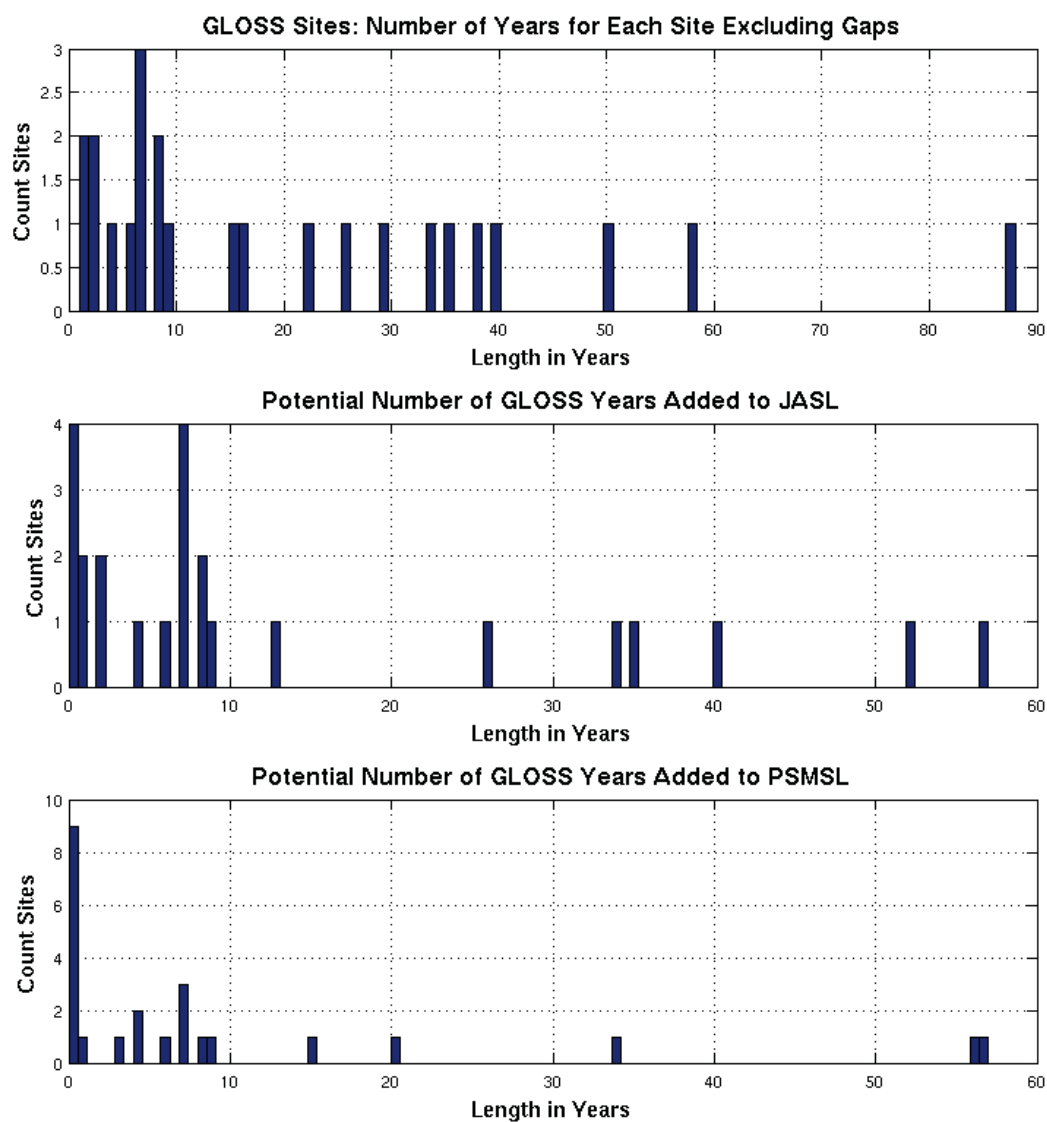


Figure 4. As for Figure 3, but exclusively for GCN sites.

Appendix A

Tide-gauge stations identified by the 2012 GLOSS questionnaire as having data at risk. The large quantity of data records in French repositories, as identified by Dr. Pouvreau, are not included.

Source	Station Name	Country	Time Span	# Yrs	Exist JASL 2012	Yrs to JASL
UKHO	Abadan	Iran	1931-1936	4		4
UKHO	Agalega Islands	Mauritius Group	1962	1		1
UKHO	Al Basrah	Iraq	1923-1930, 1932	9		9
UKHO	BARROW (RAMSDEN DOCK)	England	1849-55,74-75,82-91	17		17
UKHO	Barrow (Halfway Shoal)	England	1992-1996	5		5
UKHO	Barrow (Roa Island)	England	1992-1996	5		5
UKHO	BELFAST	Northern Ireland	1992-1993	2		2
UKHO	Belize City	Belize	?			
UKHO	Blacktoft	England	1991-1992	2		2
UKHO	BONNY TOWN	Nigeria	1963-1967	5		5
UKHO	Bournemouth	England	1974-1990	17		17
UKHO	Burnham-On-Crouch	England	1987, 1988	2		2
UKHO	Calabar	Nigeria	1961-1970	10		10
UKHO	Cascais	Portugal	1905	1	1959-2005	1
UKHO	Castries	Windward Islands	?			
UKHO	Chatham (Lock Approaches)	England	1968-74,76-79,80-87	19		19
UKHO	Coryton	England	1989-1996	8		8
UKHO	DOVER	England	1976-1985	10		10
UKHO	DUBLIN (NORTH WALL)	Ireland	1991-1993	3		3
UKHO	Dunbar	Scotland	1969-1979	11		11
UKHO	FISHGUARD	Wales	1983-1984,1986	3		3
UKHO	Fleetwood	England	1992	1		1
UKHO	GIBRALTAR	Gibraltar	1961-1999	38	1961-2000	0
UKHO	Goole	England	1994-1996	3		3
UKHO	Gorleston-On-Sea	England	1991-1993	3		3
UKHO	Gourock	Scotland	1966.68-85	19		19
UKHO	GREENOCK	Scotland	1972-84,86-89	17		17
UKHO	Haws Point	England	1982	1		1
UKHO	Heysham	England	1986-88	3		3
UKHO	HOLYHEAD	Wales	1979-88	10		10
UKHO	Honiara	Solomon Islands	1957-61,1965-68	9	1974-2009	9
UKHO	Humber Bridge	England	1991-1992	2		2
UKHO	Ilfracombe	England	1970-71	2		2
UKHO	IMMINGHAM	England	1991	1		1
UKHO	INVERGORDON	Scotland	1915-18,59-67,69-86	21		21
UKHO	Inverness	Scotland	1995	1		1
UKHO	Jabal Az Zannah	United Arab	1968-1980	13		13

		Emirates				
UKHO	Lagos Bar	Nigeria	1940-49,51-53,69-70	15	1961-70,1990-96	13
UKHO	LARNE	Northern Ireland	1968 - 1969	2		2
UKHO	LE HAVRE	France	Various			
UKHO	LEITH	Scotland	1980-1988	9		9
UKHO	Liverpool (Alfred Dock)	England	1992-1993	2		2
UKHO	MILFORD HAVEN	Wales	1980-84	5		5
UKHO	Millport	Scotland	1987-1998	12		12
UKHO	MINA AZ ZAWR (MINA SAUD)	Kuwait	1966-1967	2		2
UKHO	Mtwara Bay	Tanzania	1954-1962	9		9
UKHO	Nab Tower	England	1934-1935	2		2
UKHO	North Woolwich	England	1989-1996	8		8
UKHO	OBAN	Scotland	1910-13,1970-72	7		7
UKHO	Ogidigbe	Nigeria	1961-1962	2		2
UKHO	Oostende	Belgium	1918,1940-44	6		6
UKHO	PLYMOUTH (DEVONPORT)	England	1953-1998	46		46
UKHO	POOLE HARBOUR	England	1957-1961	3		3
UKHO	Port Harcourt	Nigeria	1963-1967	5		5
UKHO	PORT VICTORIA	Seychelles	1962-1968	7	1977-82, 1986-92	7
UKHO	PORT VILA	Vanuatu	1967-1968	3	1977-82, 1993-2009	3
UKHO	PORTLAND	England	1923-28,73-84, 85,87	20		20
UKHO	PORTSMOUTH	England	1936-37,39-58,61-96	58		58
UKHO	Ramsgate	England	1990-1991	3		3
UKHO	REYKJAVIK	Iceland	1956-1963	8	1984-1999	8
UKHO	ROSYTH	Scotland	1912-20,1945-89	40		40
UKHO	Sapele	Nigeria	1962-1969	8		8
UKHO	Scarborough	England	1958-1967	10		10
UKHO	Scrabster	Scotland	1966-1976	11		11
UKHO	SHEERNESS	England	1978-1987	10		10
UKHO	SHOREHAM	England	1965-70,88-97	16		16
UKHO	St. Mary's	England	1987-1988	2		2
UKHO	ST. PETER PORT	Channel Islands	1989-1995	7		7
UKHO	Stromness	Scotland	1910-1912	3		3
UKHO	Tema	Ghana	1963-1964	2		2
UKHO	TILBURY	England	1991-1993	3		3
UKHO	Tobermory	Scotland	1977-1978	2		2
UKHO	ULLAPOOL	Scotland	1981-1985	5		5
UKHO	Valletta	Malta	1870,1880,1903-26	25		25
UKHO	VLISSINGEN (FLUSHING)	Netherlands	1917-1918	2		2
UKHO	Warri	Nigeria	1915-1926	12		12

UKHO	WICK	Scotland	1979-1989	11		11
UKHO	Wicklow	Ireland	1968-1969	2		2
UKHO	ZEEBRUGGE	Belgium	1940-1944,1973	5		5
SHN	Puerto Belgrano	Argentina	unconfirmed	10		10
TSI	Gagra, Georgia, Black Sea	Russia	1926-1985	60		
TSI	Gudauta, Georgia	Russia	1928-1960	33		
TSI	Ochamchira, Georgia	Russia	1928-1960	33		
DMI	Rodbyhavn	Denmark	1955-1980	26		26
DMI	Kobehavn	Denmark	1890-1974	85		85
DMI	Korsor	Denmark	1890-1972	83		83
DMI	Slipshavn	Denmark	1890-1971	82		82
DMI	Flynshavn/Mommark	Denmark	1923-1969	20		20
DMI	Fredericia	Denmark	1890-1972	83		83
DMI	Aarhus	Denmark	1890-1970	81		81
DMI	Frederikshavn	Denmark	1890-1970	81		81
DMI	Hirtshals	Denmark	1890-1971	82		82
DMI	Hanstholm	Denmark	1950-1968	19		19
RWS	Hoek van Holland	Netherlands	1911-1931	21		21
IGN	Alicante I	Spain	1870-1874,1874- 1924	55		55
IGN	Almeria	Spain	1977-1998	22		22
IGN	Cartagena	Spain	1927-28,1977-89	15		15
IGN	A Coruna	Spain	1950-1983	34	1943-2008	0
IGN	Santander	Spain	1876-1924,20-28, 62-73	70		0
IGN	Tenerife	Spain	1926-1975	51	1992-2009	51
FRC	Eastport, ME	USA	1860-64,1918	6	1929-2011	6
FRC	Portland, ME	USA	1852-53,64- 66,1910-11	7	1910-2011	5
FRC	Pulpit Harbor, ME	USA	1870-1888	19		19
FRC	Portsmouth, NH	USA	1926-1935	10		10
FRC	Boston, MA	USA	1847-77,1903-11	39	1921-2011	39
FRC	Newport, RI	USA	1844-46,92-95	7	1930-2011	7
FRC	Providence, RI	USA	1872-92	21		21
FRC	Fort Hamilton, NY	USA	1893-1936	44		44
FRC	Governor's Is., NY	USA	1837-1886	50		50
FRC	Willet's Point, NY	USA	1885,1890-96	8		8
FRC	The Battery, NYC, NY	USA	1920-1935	16	1958-2011	16
FRC	Sandy Hook, NJ	USA	1835-1939	105		105
FRC	Atlantic City, NJ	USA	1911-1939	29	1911-2011	0
FRC	Philadelphia, PA	USA	1890-1937	47		0
FRC	Baltimore, MD	USA	1845,53-56,63,66, 76,86,98-99	11		0
FRC	Annapolis, MD	USA	1844-47,53,70	6		0
FRC	Old Point Comfort, VA	USA	1844-79,1906-10, 1918-19	42		42
FRC	Wilmington, NC	USA	1882,87,90-91, 1908-11	8	1935-2011	8

FRC	Charleston, SC	USA	1850-61,82-1908, 10,13	31	1921-2011	31
FRC	Fort Pulaski, GA	USA	1851-52,89-92	6	1935-2011	6
FRC	Fernandina, FL	USA	1855-61,69-71,78- 79	12	1897-2011	12
FRC	Mayport, FL	USA	1895-1939	45	1928-2000	33
FRC	Key West, FL	USA	1847,50-52,57-59, 1903	8	1913-2011	8
FRC	Cedar Keys, FL	USA	1858-60,92-93	5		5
FRC	Pensacola	USA	1890-1939	50	1923-2011	34
FRC	Biloxi, MS	USA	1855-1920	66		66
FRC	Fort Morgan, AL	USA	1846-1920	75		75
FRC	Galveston, TX	USA	1852-1939	88	1904-2011	52
FRC	San Diego, CA	USA	1853-1872	20	1906-2011	20
FRC	La Jolla, CA	USA	1924-1939	16	1924-2011	0
FRC	Long Beach, CA	USA	1924-1934	11		11
FRC	Los Angeles, CA	USA	1852-1937	86	1923-2011	70
FRC	San Francisco, CA	USA	1853	1	1897-2011	1
FRC	Sausalito, CA	USA	1851-1897	47		47
FRC	Astoria, Tongue Pt, OR	USA	1853-1876	24	1925-2011	24
FRC	Astoria, Youngs Bay, OR	USA	1931-1943	13		0
FRC	Port Townsend, WA	USA	1855,1873-77, 1933-5,41,52	11		11
FRC	Seattle, WA	USA	1891-92	2		2
NARA	Fort Sumter, SC	USA	1882-1902,04-08, 10,13	28		28
NARA	Presidio, San Francisco, CA	USA	1858,1871,1897- 1925	52		0
NARA	Astoria	USA	1853-1858,60,76, 1925	9	1925-2011	8
NARA	Olympia, Puget Sd., WA	USA	1916-1924	9		9
NARA	Craig, Prince of Wales Is, AK	USA	1914-1918,1920	6		6
NARA	Ketchikan, AK	USA	1911,1914-15,18- 25	11	1918-2011	3
NARA	Kodiak, AK	USA	1880-91,1906-9, 18-20,32-39,49-74	48	1975-2010	48
NARA	St. Paul, Kodiak, AK	USA	1880-1891	12		12
NARA	Sitka, AK	USA	1893-97,1924-25	7	1938-2011	7
NARA	San Juan, PR	USA	1892-1897,99	7	1977-2011	7
NARA	St. Thomas, USVI	USA	1872-75,1923-25	7		7
NARA	Manila	Philippines	1901-1940	40	1984-2008	40
NARA	Cebu	Philippines	1935-1938	4	1998-2008	4
NARA	Miami Beach, FL	USA	1931-1938	8		8
NARA	Mobile, AL	USA	1934-1937	4		4
NARA	Long Beach, CA	USA	1935-1936	2		2
NARA	Santa Barbara, CA	USA	1931-1935	4	1996-2011	4

NARA	Santa Monica, CA	USA	1933-1938	6	1973-2011	6
NARA	Friday Harbor, WA	USA	1934-1938	5		5
NARA	Olympia, Puget Sd., WA	USA	1934-1935	2		2
NARA	Toke Pt, Willapa Bay, WA	USA	1935-1938	4	1972-2011	4
NARA	Truk, Caroline Is.	Fed. St. Micronesia	1948-1949	2	1963-1991	2
NARA	Honolulu, HI	USA	1883-84,92-99, 1901-4	14	1877-92, 1905-2011	7
LINZ	Wellington	New Zealand	1887-1944	58	1944-2010	57
LINZ	Lyttleton	New Zealand	1883-1923	40	1994-2010	40
LINZ	Dunedin	New Zealand	1883-1899	15	1985-2010	15
LINZ	Auckland	New Zealand	1899-1902	4	1984-1988	4
LINZ	West Port	New Zealand	1901-1981	80	1984-1985	80
CHS	Tofino, BC	Canada	1905-1908,1917-1948	35	1963-2010	35
CHS	Point Atkinson, BC	Canada	1922-1961	40		40
CHS	Vancouver, BC	Canada	1901;05-08,1924-1942	23		23
CHS	Prince Rupert, BC	Canada	1906-08, 1919-1942	26	1910-18; 1963-2010	26
CHS	Port Hardy, BC	Canada	1905-1909	5		5
CHS	Victoria, BC	Canada	1899-1905	7	1909-2007	7

Environmental Data Through Time

Extending The Climate Record

Stephen Del Greco

Chief Climate Services and Monitoring Division, National Climatic Data Center (NCDC), National Oceanic and Atmospheric Administration (NOAA), USA, Stephen.A.Delgreco@noaa.gov

Abstract

The National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center (NCDC) is responsible for acquisition, archive, and dissemination services for climate and environmental data and information that fulfill much of the Nation's climate data requirements. Those include stewardship for in situ, satellite and radar data and information. Over 5 petabytes of data reside in the archives, and growth trends over the next several years are expected to increase tenfold. The Center is also assigned the analytic role of describing the climate and providing scientific assessments, such as the Climate Change Science Program (CCSP), "State of the Climate" reports and National and Global Assessments. Towards that end, NCDC has several programs that involve extending the climate record, and the NCDC led the Climate Database Modernization Program (CDMP). CDMP provided substantial funding between 1999 and 2012 to rescue and preserve historic climate and environmental data. CDMP has currently placed online over 57 million images and some 14 terabytes of weather and environmental data. In addition, hourly weather records keyed through CDMP continue to be integrated into NCDC's digital database, extending the period of record for many stations back into the 1800s.¹ Millions of data images available online also have associated environmental data that need to be digitized. NCDC is using its remaining CDMP resources and Crowdsourcing to digitize and analyse those climate data. In partnership with the Cooperative Institute for Climate and Satellites (CICS), it is deriving Climate Data Records (CDRs) for atmosphere and terrestrial features using satellite data (Global Essential Climate Variables) that date back to the 1970s. The Center also performs research in Paleoclimatology. Paleoclimate data come from natural sources such as tree rings, ice cores, corals, and ocean and lake sediments and extend the archive of weather and climate information back hundreds to millions of years. The Center maintains the world's largest archive of climate and paleoclimate data. While it is important to highlight the development of proxy datasets using paleo data and the development of satellite-based CDRs, this paper focuses on extending the climate record by rescuing past observational data.

Author

Stephen Del Greco is the acting Chief of NOAA's National Climatic Data Center (NCDC) Climate Services and Monitoring Division (CSMD). He provides oversight, leadership and management for CSMD activities that include User Engagement and Services, Climate Monitoring and Assessments, Data Access and Applications, and Regional Climate Services. The division develops and maintains systems to meet NOAA planning and execution for data discovery, data dissemination, data display and data interoperability. The division also delivers climate products to operations in support of assessment of the State of the Climate Regionally, Nationally and Internationally. Stephen's primary focus is providing climate services and robust access capabilities for weather and climate information using data federated systems and tiered services. He holds a degree in Atmospheric Sciences from the University of North Carolina and attended Harvard University, John F. Kennedy School of Government, Public Service Executive Education program.

¹ Angel, W, T. F. Ross, and R. T. Truesdell, "NOAA's Climate Data Modernization Program Paving the Way for Data Stewardship," in *91st AMS Annual Meeting, 22-27 January 2011, New Orleans, LA, 27th Conference IIPS [International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology]*, American Meteorological Society, Boston, Mass., File 4B.4 (January 2011).

1. NOAA partnerships for data rescue - U.S. data

In partnership with the private industry, the NOAA's NCDC imaged and keyed over 56 million images and over fourteen terabytes of data from paper and microfilm records. Those data are available free of charge via the Image and Publication System for images (IPS)² and the Climate Data Online (CDO) System for digital records.³ While NCDC continues to image and digitize retrospective weather and climate data using remaining CDMP resources, the center is partnering with other institutions and also transitioning to using public volunteers to digitize records through crowdsourcing⁴ and Citizen Science Alliance⁵ programs. Several completed and/or ongoing projects and partners are listed and briefly described below:

1.1 Surface Airways Observations (SAO) Project

NOAA contributed to extending the U.S. climate record by rescuing National Weather Bureau and National Weather Service (NWS) Surface Airways Observations (SAO) data for NWS sites dating back to 1893. The project imaged over 2 million forms and keyed over 400 million hourly surface observations records and added over 50 years of hourly/synoptic data for over 700 U.S. locations [figure 1]. The SWO library is the largest in IPS.

1.2 U.S. Founding Fathers Weather Journals

Weather and climate data recorded by George Washington, Thomas Jefferson and Benjamin Franklin and other colonists are archived in original manuscripts, microfilmed and stored at the National Archive and Records Administration (NARA). The records were also imaged and are available on IPS. These colonial diaries and data are a treasure trove for the climatologist seeking data on climate of the 18th and 19th century [figure 2].

1.3 U.S. Forts Project

NOAA partnered with the Midwest Regional Climate Center⁶ to image and digitize historical climate data from U.S. Army forts. The "Forts Project", focused on imaging and keying data from 1820-1892 for Army forts; however, other sources of climate data, such as Smithsonian Institution's 19th century network of voluntary observers, United States Signal Service observations and private citizen observation journals, were included in the program. The digitized data went through extensive quality control processes prior to becoming available to the public; the forms [figure 3], some almost 200 years old, are available on IPS, and the digitized data on CDO.

1.4 Shoreline Vectorization - National Ocean Service/Coastal Services Center

A digital national shoreline database used for spatial analysis of coastal areas. Rescue work converts topographic sheet images to a geo-referenced vector format. These shoreline data are used to help protect

² National Climatic Data Center Image and Publication System, <http://www7.ncdc.noaa.gov/IPS/>.

³ NOAA National Climatic Data Center Climate Data Online, <http://www.ncdc.noaa.gov/cdo-web/webservices>.

⁴ Crowdsourcing web page, <http://www.crowdsourcing.org/>.

⁵ Citizen Science Alliance, <http://www.citizensciencealliance.org/>.

⁶ Midwest Regional Climate Center, <http://mcc.sws.uiuc.edu/>.

coastal resources, sustain the environmental quality of the coastal environment, and mitigate impacts from coastal processes.⁷

1.5 Defense Meteorological Satellite Program (DMSP) Film Imaging - National Geophysical Data Center

Scanning of DMSP film from the early 1970s to the mid-1980s. The DMSP film contains observations relevant to global cloud climatology, hurricane and typhoon climatology, the extent and conditions of polar ice, continental and mountain snowpack, and the record of expansion in human settlements.⁸

1.6 Imaging of NOAA Central Library Holdings - NOAA Central Library

Imaging of foreign climate data books; U.S. daily weather maps from 1871 to the late 1960s; and, in coordination with the American Meteorological Society, imaging of the Monthly Weather Reviews. These images are made available online at the NOAA Central Library.⁹

1.7 Imaging of Historical U.S. Coast Pilot Editions - National Ocean Service/Office of Coast Survey

The Coast Pilot collection consists of approximately 800 volumes from the 1800s to today. The volumes are available online at the Office of Coast Survey and the NOAA Central Library. The collection includes significant navigation information, and descriptions and locations of localized atmospheric features and conditions.¹⁰

1.8 Digitization of Ionospheric Data - National Geophysical Data Center

Digitization of ionosphere bottom side vertical incidence sounding data values from the 1930s through 1957. The paper media contain both half hourly and hourly data.¹¹

1.9 Digitization of U.S. Upper Air Pilot Balloon (Pibal) Data - National Weather Service

Imaged (from film) and keyed U.S. pibal observations prior to 1948.

2. NOAA partnerships for data rescue - International Projects

NOAA partnered with 27 countries across several continents to rescue surface, marine and upper air data. For example NOAA partnered with the World Meteorological Organization¹² to rescue weather balloon upper air data in seven African nations: Kenya, Malawi, Mozambique, Niger, Senegal, Tanzania and

⁷ Shoreline Mapping, <http://shoreline.noaa.gov/>.

⁸ Defense Meteorological Satellite Program (DMSP) Film Imaging, <http://www.ngdc.noaa.gov/dmsp/index.html>.

⁹ NOAA Central Library, http://docs.lib.noaa.gov/rescue/data_rescue_home.html.

¹⁰ Imaging of Historical U.S. Coast Pilot Editions, <http://www.nauticalcharts.noaa.gov/nsd/hcp.htm>.

¹¹ Digitization of Ionospheric Data - National Geophysical Data Center, <http://www.ngdc.noaa.gov/stp/IONO/ionohome.html>.

¹² World Meteorological Organization, http://www.wmo.int/pages/index_en.html.

Zambia [figure 5]. The African countries imaged their data locally, and sent the images to NOAA's NCDC for keying and uploading to IPS and NOAA's Integrated Global Radiosonde Archive Database (IGRA).¹³ IGRA consists of radiosonde and pilot balloon observations at over 1500 globally distributed stations [figure 6]. Observations are available for standard surface, tropopause and significant levels. Variables include pressure, temperature, geopotential height, dew point depression, wind direction and wind speed. Over 150,000 images of pibal (upper air wind) records from the 1940s to 2003 from the 7 African countries are digitized. The digital data files were also provided to the host countries that imaged the data while the keyed data files are hyperlinked to the actual images, providing an easy access to the original records. Another data rescue example is the partnership with the United Kingdom to image and digitize the marine logbooks in the British Archives. It included the English East India Company (EIC) Instrumental Observations 1789-1834. The Met Office Hadley Center imaged over 1100 of the original 2000 logbooks of the English East India Company (EIC) held at the British Library. The selection of logs was based on their holdings of weather observations, visual and instrumental, as well as the significant spatial coverage of voyages from England to India and China through the Atlantic, Indian and Southern Oceans during those years. The EIC collected commenced well before the landmark 1853 Brussels Maritime Conference¹⁴ devoted to coordinating an international effort for global systematic collections of marine instrumental and visual observations, and is probably the largest and earliest collection of such systematic instrumental observations. The project captured all noon observations containing location, instrumental observations of pressure and air temperature (and occasionally sea surface temperature), and visual estimates of winds, state of weather and state of sea. From the digitized logbooks, over 285K observations were digitized, significantly increasing early instrumental coverage both spatially and temporally.¹⁵

2.1 Crowdsourcing

Rescue of United Kingdom marine data continues. Weather observations made by Royal Navy ships around the time of World War I were recently digitized as part of a Zooniverse (crowdsourcing)¹⁶ project called oldWeather.¹⁷ Volunteers digitized over 1 million, six hundred thousand Royal Navy-derived weather observations. That was the first phase for the project; oldWeather is currently in phase two, to digitize weather observations by ships in the Arctic. The citizen volunteers have completed 46% of the ship logs for recovering Arctic and worldwide weather observations made by United States ships since the mid-19th century.

3. Paleoclimatology – Extending the climate record using “proxies”

Paleoclimatology is the study of past climate prior to instrumental weather measurements. Paleoclimatologists use information from natural climate “proxies,” such as tree rings, ice cores, corals,

¹³ Integrated Global Radiosonde Archive Database (IGRA), <http://www.ncdc.noaa.gov/oa/climate/igra/index.php>.

¹⁴ Maury, M. F., “Maritime Conference held at Brussels for devising a uniform system of meteorological observations at sea, August and September, 1853,” in *Explanations and Sailing Directions to Accompany the Wind and Current Charts*, 6th ed. (E. C. and J. Biddle: Philadelphia, 1853), 54-96.

¹⁵ Woodruff, S., S. Worley, S. Lubker, Z. Ji, E. Freeman, D. Berry, P. Brohan, E. Kent, D. Reynolds, S. Smith, and C. Wilkinson, “ICOADS Release 2.5: Extensions and Enhancements to the Surface Marine Meteorological Archive,” *International Journal of Climatology* 31, no. 7 (2011): 951-967.

¹⁶ Zooniverse web page, <https://www.zooniverse.org/>.

¹⁷ oldWeather web page, <http://www.oldweather.org/>.

and ocean and lake sediments, that record variations in past climate [figures 7, 8]. Records of past climate from such proxy records are important for several reasons. Instrumental records of climate are limited in many parts of the world to the past 100 years or less, and are too short to assess whether climate variability, events, and trends of the 20th and 21st centuries are representative of the long-term natural variability of past centuries and millennia. For example, was the 1930s Dust Bowl drought, a widespread and severe event in the United States, a rare occurrence or have similar events occurred in past centuries? Knowledge of the long-term natural variability of the Earth Climate system, and its causes, will also allow an understanding of the roles of natural climate variability and human-induced climate change in the current and future climate. In particular, reconstructed temperatures from proxy data for the past 1000 years have allowed an assessment of the warming over recent decades.¹⁸

4. Climate Data Records – Extending the climate record using satellite data

NOAA's NCDC recently initiated a satellite Climate Data Record (CDR) program to provide continuously objective climate information derived from weather satellite data that NOAA has collected for more than 30 years. Those data comprise the longest record of global satellite mapping measurements in the world, and are complemented by data from other sources including NASA and Department of Defense satellites and foreign satellites. The mission of NOAA's Climate Data Record Program is to develop and implement a robust, sustainable, and scientifically defensible approach to producing and preserving climate records from satellite data [figure 9]. For the first time, NOAA is applying modern data analysis methods, which have advanced significantly in the last decade, to these historical global satellite data. The program will unravel the underlying climate trend and variability information and return new economic and scientific value from the records. In parallel, NCDC will maintain and extend these Climate Data Records by applying the same methods to present-day and future satellite measurements. The results will provide trustworthy information on how, where and to what extent the land, oceans, atmosphere and ice sheets are changing. In turn, this information will be used by energy, water resources, agriculture, human health, national security, coastal community and other interest groups. The CDR data will improve the Nation's resilience to climate change and variability, maintain our economic vitality, and improve the security and well-being of the public.¹⁹

5. Conclusion

OldWeather project's mission statement, "Old Weather: Our Weather's Past, the Climate's Future" is one that resonates with climatologists. To gain better understanding of the earth's physical processes and the future state of the climate it is essential to be able to reconstruct past climates. NOAA continues to work with partners in both the public and private sectors to look for innovative ways to rescue retrospective weather and climate data from deteriorating media, and to make the data available in digital formats and accessible online for the public. Once converted into electronic formats, the data are more portable, can quickly and easily be shared, and contribute further to global climate studies.

¹⁸ NOAA National Climatic Data Center Paleoclimatology web page, <http://www.ncdc.noaa.gov/paleo/paleo.html>.

¹⁹ Climate Data Records, <http://www.ncdc.noaa.gov/cdr/index.html>.

Figures

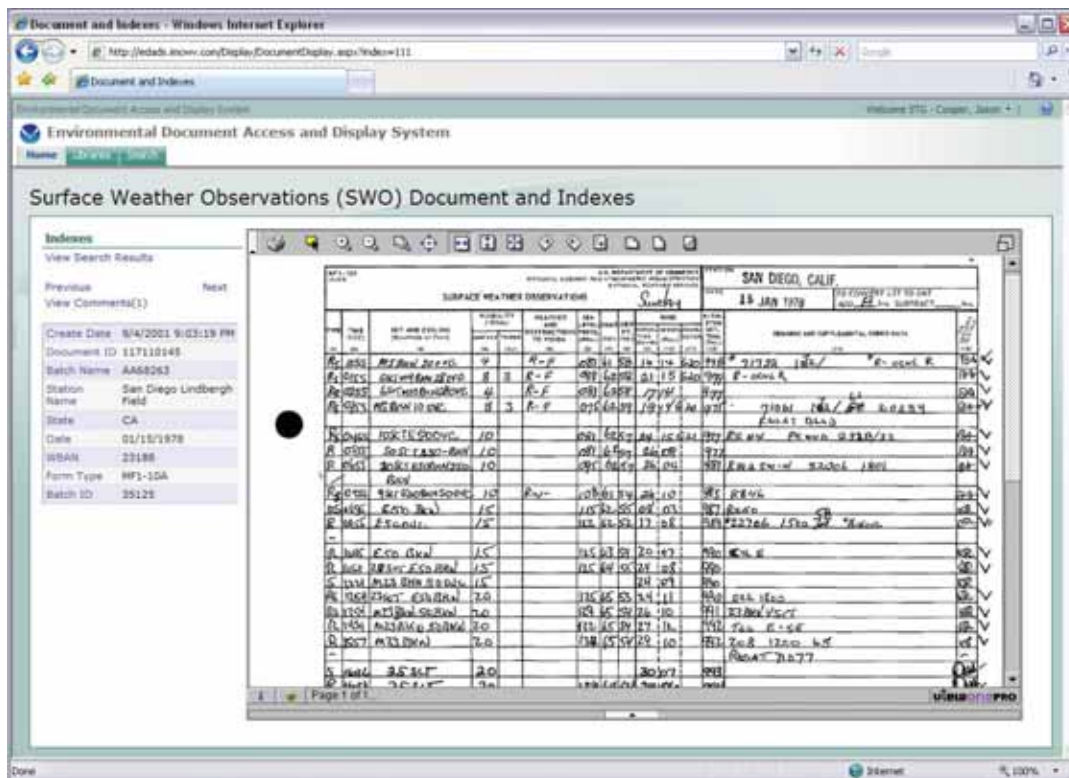


Figure 1. SAO Surface Weather Observation image.

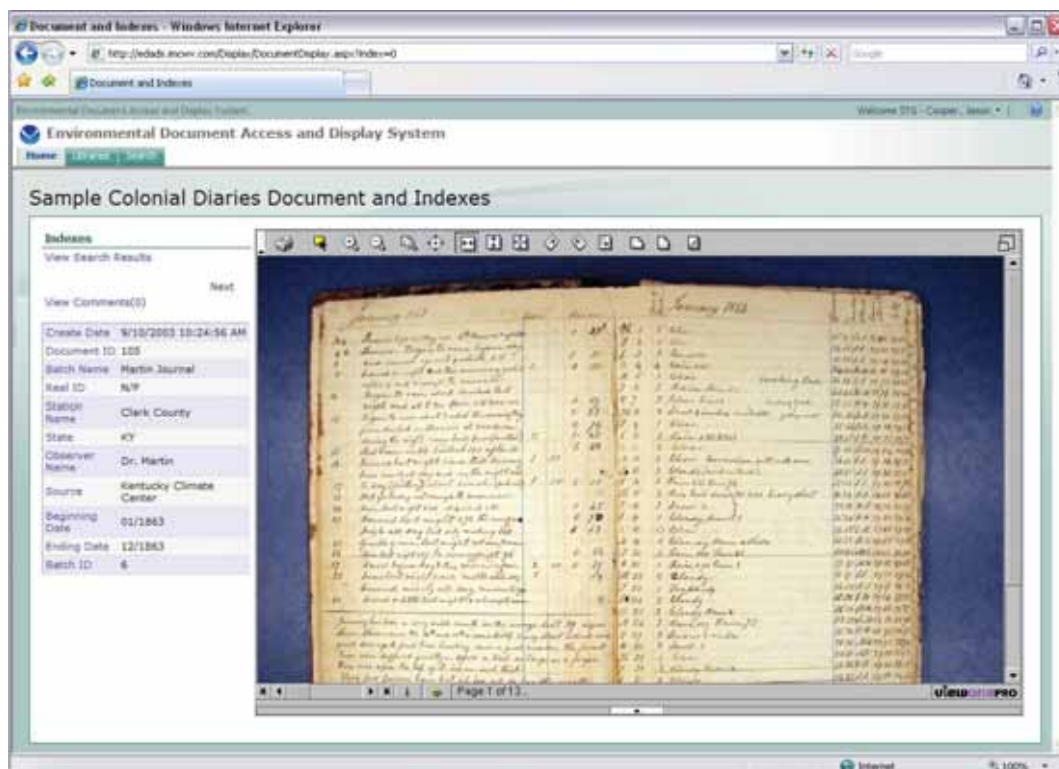


Figure 2. Colonial Log Book image.

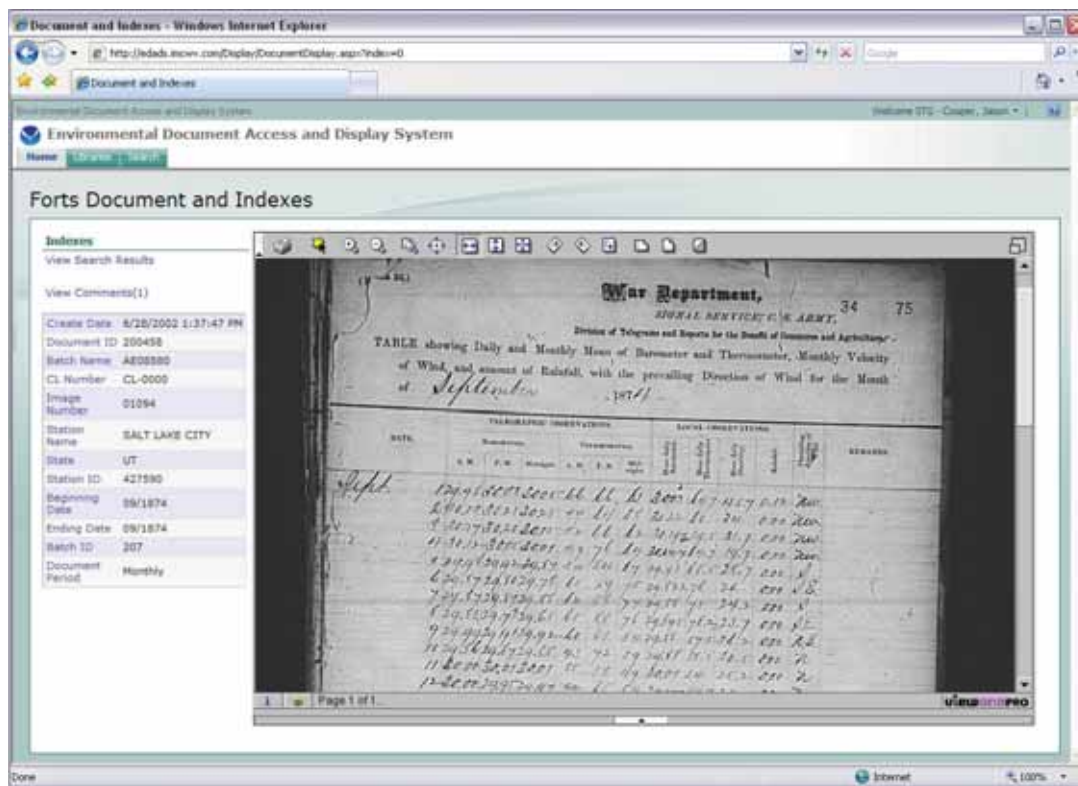


Figure 3. Forts Weather Observations image.

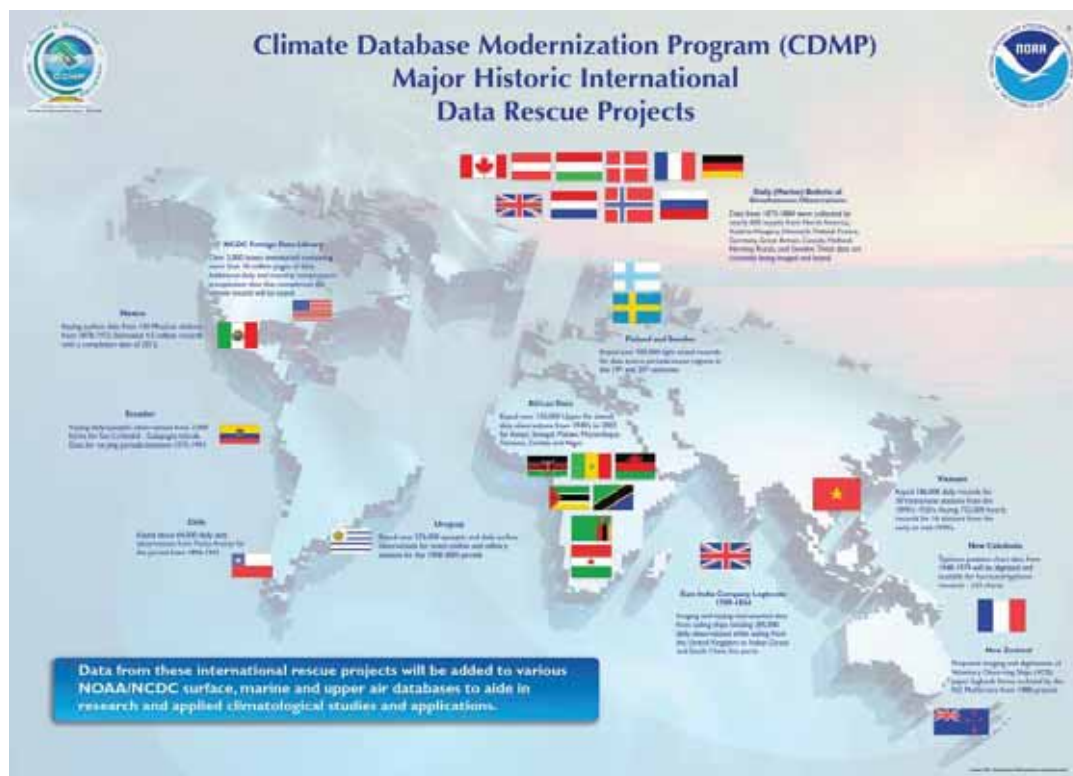


Figure 4. International data rescue projects.

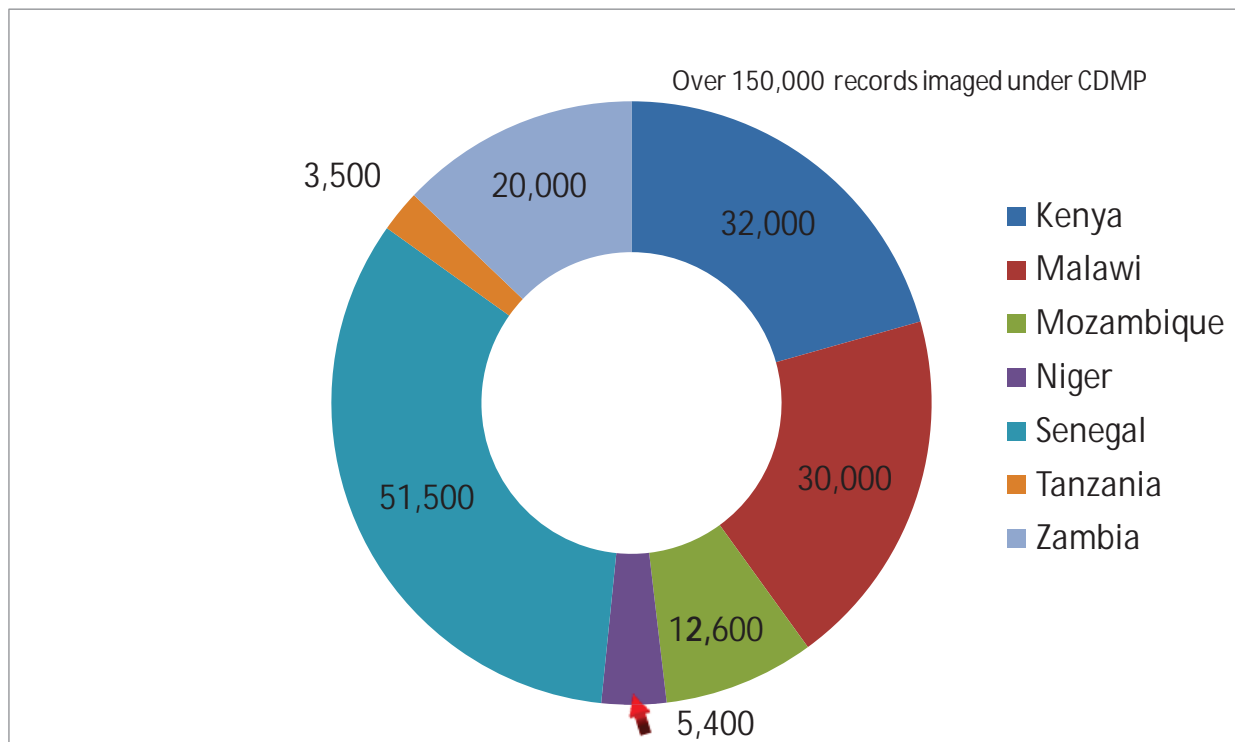


Figure 5. Upper Air data rescued in Africa.

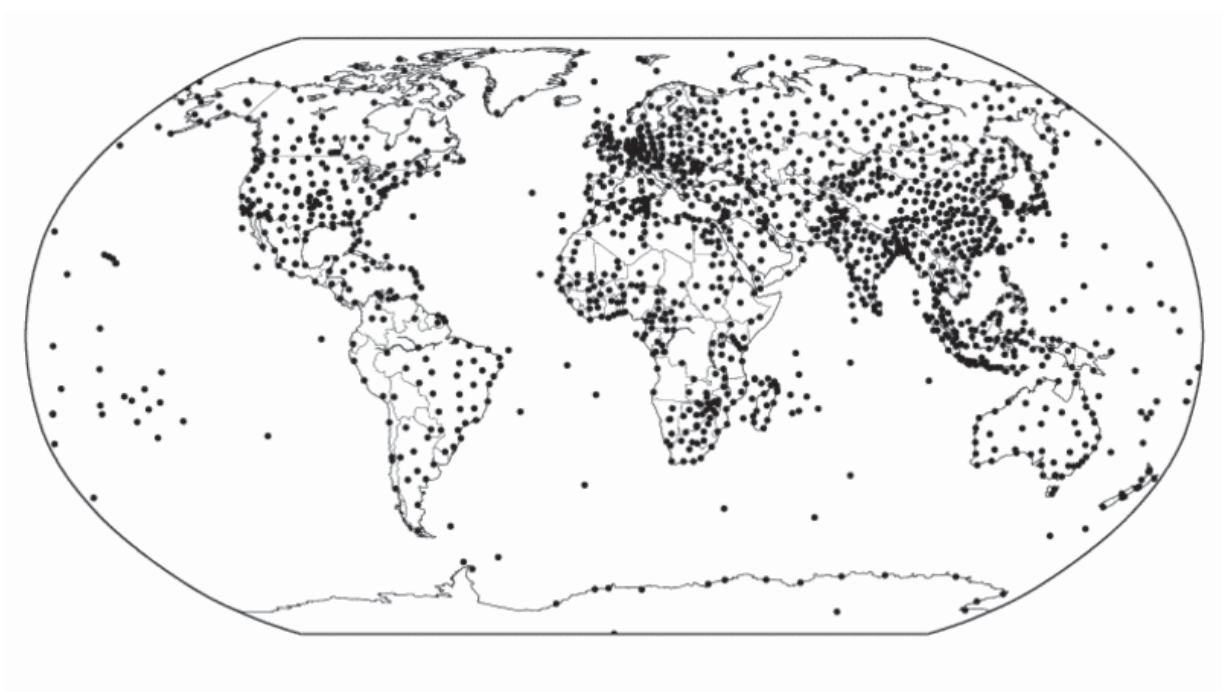


Figure 6. Locations of all stations in IGRA.

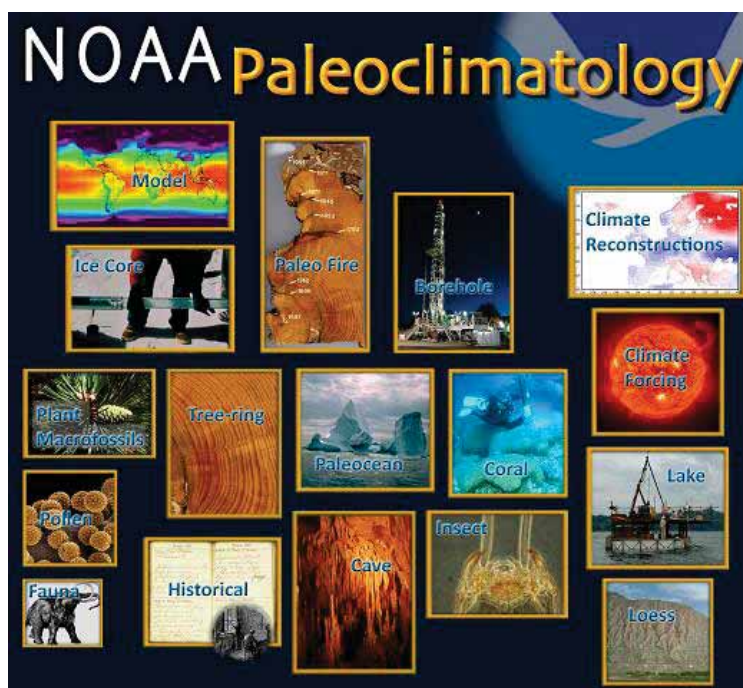


Figure 7. Types of proxy datasets used in describing past climates.

The interval of time spanned by different paleoclimate proxies, displayed on a logarithmic scale.

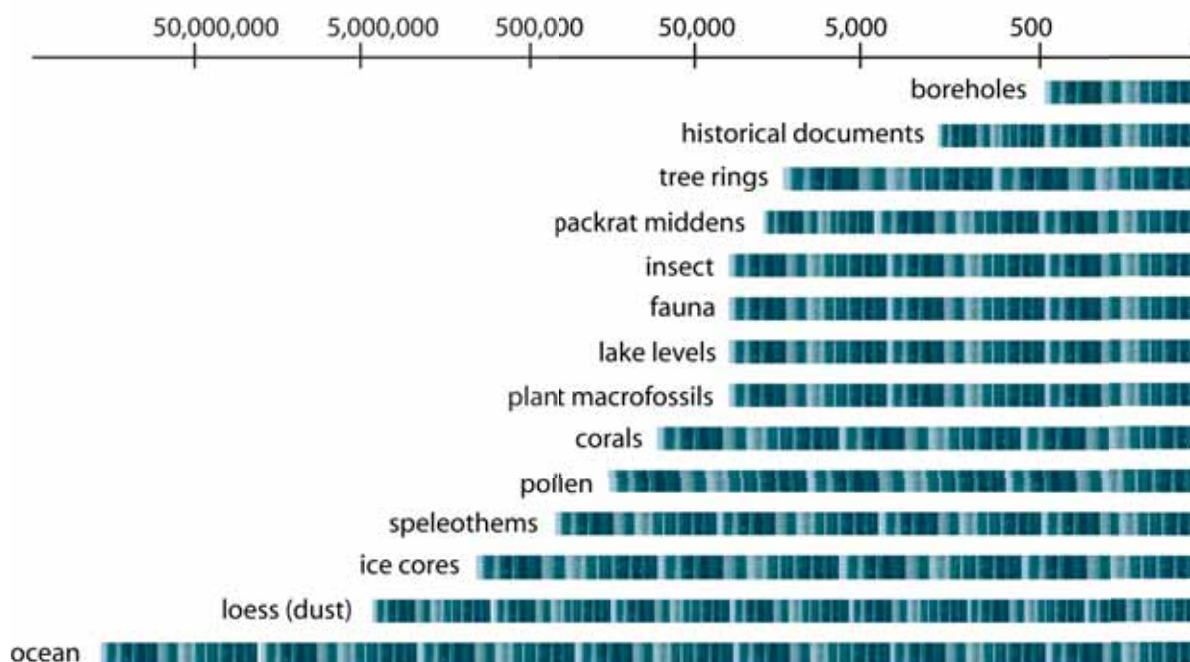


Figure 8. Paleoclimate proxies time scales.

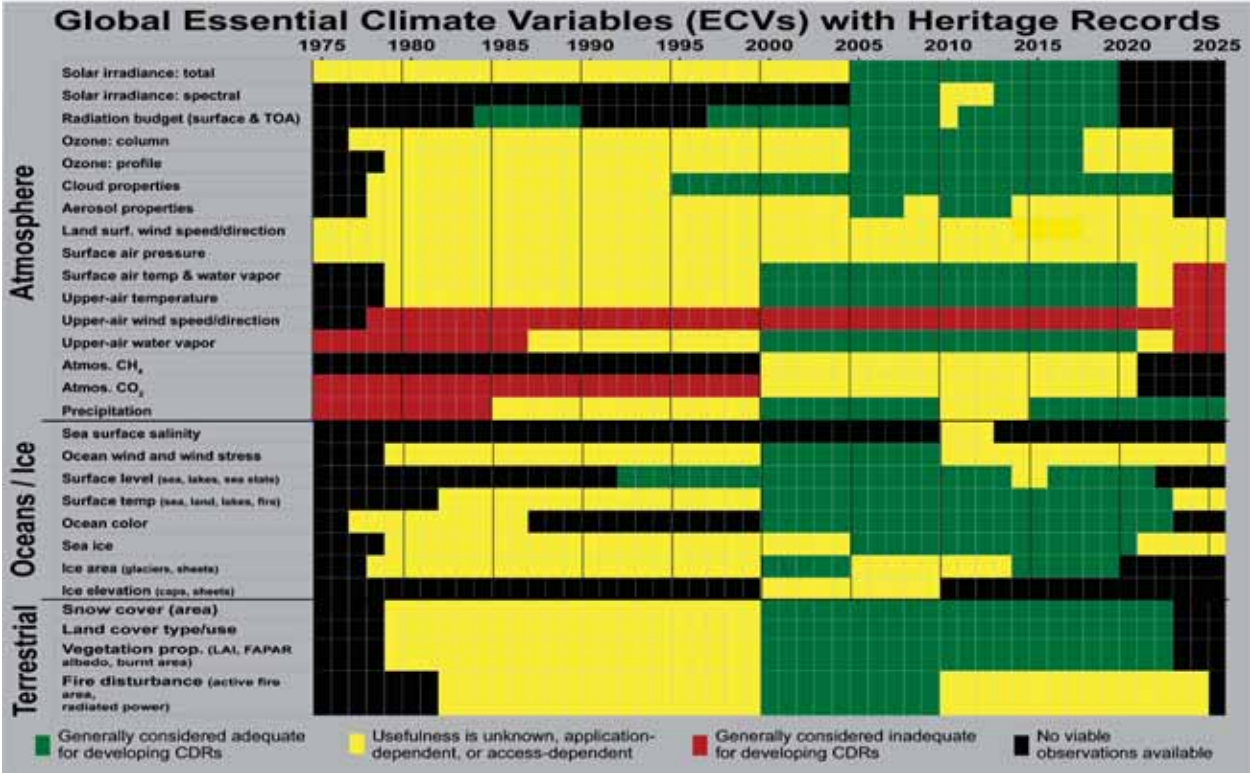


Figure 9. Climate Data Records Essential Climate Variables.

The Map as a Fundamental Source in the Memory of the World

Tracey P. Lauriault¹ and D. R. Fraser Taylor²

¹Postdoctoral Fellow, Geomatics and Cartographic Research Centre (GCRC), Carleton University, Canada, tlauriau@gmail.com; ²FRSC, Distinguished Research Professor and Director of the Geomatics and Cartographic Research Centre, Carleton University, Canada; member of the CODATA Data at Risk Task Group, fraser_taylor@carleton.ca

Abstract

Maps and spatial information have been fundamental facets of the memory of societies from all over the world for millennia, and their preservation should be an integral part of government strategies in managing digital data. The digital era in map-making is a relatively recent activity; the first digital maps date from the 1960s. Digital mapping has accelerated very rapidly over the last decade and is now ubiquitous with an increasing amount of spatially referenced information being created by non-governmental organizations, academia, the private sector and government, as well as by social networks and citizen scientists. Unfortunately, despite that explosion of digital mapping little or no attention is being paid to preservation. As a result, the very maps that have been such a fundamental source of scientific and cultural information are now seriously at risk. Already we are losing map information faster than it is being created, and the loss of that central element of the cultural heritage of societies all over the world is a serious concern. There has already been a serious loss of maps; the Canada Land Inventory and the 1986 BBC Domesday Project are only two such casualties, and mapping agencies all over the world are struggling to preserve maps in the new digital era. It is somewhat paradoxical that it is easier to get maps that are hundreds, and in some cases thousands, of years old than maps of the late 20th and early 21st centuries. This paper examines the opportunities and challenges of preserving and accessing digital maps, atlases and geospatial information, all of which are Canada's cultural and scientific knowledge assets.

Authors

Dr. Tracey P. Lauriault is a Postdoctoral Fellow at the Geomatics and Cartographic Research Centre working on the SSHRC Partnership Development Project entitled Mapping the Legal and Policy Boundaries of Digital Cartography in collaboration with the Centre for Law, Technology and Society; the Canadian Internet Public Policy Clinic and Natural Resources Canada. She also participates in activities and represents the GCRC on topics related to the access to and preservation of data. As a citizen she is an open data advocate and is the co-founder of civicaccess.ca and datalibre.ca.

Dr. Professor Taylor is a Distinguished Research Professor of International Affairs and Geography and Environmental Studies at Carleton University, Canada and is also Director of the Geomatics and Cartographic Research Centre. He is a Fellow of the Royal Society of Canada and is widely recognized as one of the world's leading cartographers. Dr. Taylor's main research interests in cartography lie in the application of geographic information processing to the analysis of socio-economic issues and the presentation of the results in the form of cybercartographic atlases. Cybercartography is an innovative new concept which he first introduced in 1997. His current funded research involves working with aboriginal and Inuit communities to empower these communities to express their perceptions of their own environmental and socio-economic reality in new ways utilizing the Cybercartographic Atlas Framework. This includes the innovative open source software called Nunaliit. Dr. Taylor is Chair of the International Steering Committee for Global Mapping (ISCGM) and the International Advisory Group on the Arctic Spatial Data Infrastructure, he is also a member of the Mapping Africa for Africans Working Group of the International Cartographic Association, a member of the Joint Board of the Geospatial Information Societies and a member of the UN initiative on Global Geospatial Information Management.

1. Introduction

Maps do more than help us locate things. Maps and atlases, like books, reports, and archival records, enable us: to understand the territorial evolution of our nations, to see how colonial surveyors drew property boundaries, to visualize the spaces negotiated in treaties, and to assess the distribution of population and national resources. They can model climate change, the economy, or display the spatial relationships between cities, infrastructure and society. Maps inform planning, policy and economic decisions, while also shaping how we imagine spaces. They delight, are repositories of cultural and scientific knowledge and are a key part of our collective geographical memories.

In Canada, all levels of government create maps. At the Federal level, Natural Resources Canada (NRCan) produces, acquires and maintains the largest collection of maps, satellite and radar imagery, air photos, geospatial datasets, and publishes the *Atlas of Canada*. These are disseminated via geospatial data portals, map servers, and special programs such as *GeoGratis*. NRCan is also home to Canada's Geospatial Data Infrastructure (CGDI), which comprises the open and interoperable standards, specifications, practices, technology, framework data, operational policies and institutional environment to disseminate Canada's geospatial data assets on the Internet. Numerous other federal departments also produce geospatial data.¹ Geospatial data also represent the largest collection of freely accessible datasets currently being disseminated in the Treasury Board Secretariat's (TBS) Open Data Pilot portal.² Provincial and territorial governments also produce maps, are responsible for cadastres, and have geospatial databases to help manage transportation networks, watersheds, natural resources and to administer programs such as health and education. At the level of the city geographic information system (GIS) units can be found within information technology (IT) sections to manage infrastructure and other municipal responsibilities. Academic institutions create maps and atlases which are normally funded by government granting agencies, foundations or special programs such as the International Polar Year (IPY). Canadian non-governmental organizations (NGOs) produce maps on a variety of topics such as food security, water quality and population health. The private sector through companies such as Google, Autodesk, and ESRI is a major producer of digital maps and geomatics products in Canada and is a multibillion-dollar industry.³ The public and communities are also engaged in map-making endeavours by contributing volunteered geographic information (VGI)⁴ into Mashups,⁵ building local infrastructures,⁶

¹ For example the Public Health Agency of Canada; Fisheries and Oceans; Environment Canada; Aboriginal Affairs and Northern Development Canada; Citizen and Immigration and many others.

² See TBS Open Data Portal <http://www.data.gc.ca/default.asp?lang=En&n=F9B7A1E3-1> (accessed 29 August 2012).

³ See Statistics Canada, *Surveying and Mapping Services Bulletin*, (2010) Catalogue no. 63-254-X, available at <http://www5.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=63-254-XIE&lang=eng#formatdisp> (accessed 28 August 2012).

⁴ VGI is georeferenced user-contributed information that is collected and disseminated in such a way that others can use them. For more details see Teresa Scassa, "Legal issues with volunteered geographic information," 2012, *Canadian Geographer*, available at <http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0064.2012.00444.x/abstract> doi:10.1111/j.1541-0064.2012.00444.x.

⁵ The Regroupement activists pour l'inclusion Québec (RAPLIQ), with Montreal Ouvert and OpenNorth collaborated to create a prototype accessibility audit of commercial establishments. People with disabilities collected VGI data and these were mapped using GoogleMap's open API. The Montréal Accessible mashup is available at <http://montrealaccessible.ca/> (accessed 18 August 2012).

⁶ The Canadian Council on Social Development (CCSD) created a membership based consortium called the *Community Data Program* that includes myriad georeferenced demographic and socio-economic data available at <http://communitydata-donneescommunautaires.ca/Home>. The Social Planning Council of Ottawa has created a Community Information and Mapping System (CIMS) as seen here <http://www.cims-scic.ca/> (accessed 29 August 2012).

creating mobile device apps,⁷ and by being engaged in participatory mapping⁸ or citizen science endeavours.⁹ Digital maps and geospatial data are ubiquitous artefacts in the 21st century.

Geospatial data, maps and atlases, have been produced for millennia,¹⁰ and are cultural, historical and scientific records of knowledge and advancement, which makes them a fundamental source in memory of the world. This source of knowledge, in its borne digital form, is however not being systematically preserved. The digital era in map making is a relatively recent phenomenon of the last fifty years¹¹ and has accelerated very rapidly over the last decade as software has become easier to use and are less expensive, geospatial data are more accessible, organizations have opened their application programming interfaces (APIs) and Internet social networking services have proliferated.

Unfortunately, little, or no, attention is being paid to preserving maps, atlases and their related data whether created by authoritative or non-authoritative sources, as a result, these are very much at risk. Already we are losing spatial information faster than it is being created and the loss of this central part of the cultural heritage of societies all over the world is a serious concern.

This paper will consider seven related topics. It will begin by illustrating the problem by discussing two map and atlas rescue and salvage endeavours. Secondly, cybercartographic atlases are now mapping traditional knowledge (TK) by applying participatory research techniques. These are considered by the Indigenous communities involved repositories of local, cultural, scientific and historical knowledge. In creating these atlases preservation is considered from the outset even though an adequate archive to ingest them has still not emerged. Thirdly, while most geospatial data in Canada are not preserved, there are some geospatial data archives and these will be discussed. Fourthly, data management (including preservation) and access has become a key policy issue in Canada and a number of national public consultations on the topic have taken place. Their outcomes will be reviewed. Fifthly, there are Canadian legislation, directives and policies, which mandate the management of government information assets, and a brief overview of those related to spatial data will be examined. Sixthly, geospatial data and maps are also official government records explicitly referred to in Canadian acts and regulation and those for which the Minister of NRCan is responsible are discussed. Seventhly, the preservation of maps, atlases and geospatial data has been examined in Canadian research and some results are showcased. Finally, the paper will conclude by examining the key issues presented and will suggest strategies for the preservation of these cultural and scientific assets.

⁷ Open data initiatives make data accessible to developers who create GeoApps such as *Restonet* (<http://restonet.ca/>) which maps food inspection results; Edmonton's *Historic Buildings Android App* (<http://contest.apps4edmonton.ca/apps/21>) or the BC Apps4ClimateChange contest apps such as *Waterly* (<http://www.waterly.ca/>) which tracks rainfall and combines these with watering restrictions providing users with a watering schedule.

⁸ The Inuit Sea Ice Use and Occupancy Project (ISIUOP), which produced the Inuit siku (sea ice), *Atlas* (<http://sikuatlas.ca/index.html>) which include data contributed and mapped by Inuit hunters and elders.

⁹ The *Water and Environmental Hub* (WeHub) project “is a cloud-based, open source web platform that aggregates, federates, and connects water data and information with users looking to download, analyse, model and interpret water and environmental-based information”. Citizens can upload their data to the system available at <http://www.watereenvironmentalhub.ca/about/project>. Also see Digital Fishers where citizens load observations into a crowdsourced database and analyse results, available at <http://digitalfishers.net/> (accessed 18 August 2012).

¹⁰ For example: the Bedolina and Giadighe petroglyphs at Valcamonica (2500 BC), Çatal Hüyük (6200 BC) wall paintings, the Hecataeus of Miletus *Geographica* maps (c. 550 – 476 BCE), and Ptolemy's World Map (circa 150).

¹¹ The Canada Geographic Information System (CGIS) developed in 1960s is recognized as the first operational GIS. Dr. Roger Tomlinson who was then based Department of Forestry and Rural Development developed it.

2. The Rescue and Salvage of the Canada Land Inventory (CLI) and the 1986 Domesday Project

The Canada Land Inventory (CLI) was initiated in 1960 by the Department of Forestry and Rural Development. In 1963 the Canadian Geographic Information System (CGIS) was developed to automate the data management and mapping of CLI data. The CGIS was the world's first GIS, it was a milestone in the history of geographic and government computerization and it revolutionized mapping. One of the principal driving forces behind the CGIS "was the idea that the CLI maps could be interpreted and analysed in a myriad of ways if the information could be manipulated by computers."¹² It was established "as a joint federal-provincial project to guide the development of policy on the control and management of land-based resources."¹³ The CGIS ultimately grew to contain thousands of maps and unknowingly became a technology that "spawned an industry that today is worth billions of dollars."¹⁴ The CLI was conceived to "classify lands as to their capabilities; to obtain a firm estimate of the extent and location of each land class and to encourage use of CLI data in planning."¹⁵

The CLI was an incredibly ambitious program that mapped 2.6 million square kilometers of Canada and "the original cost of the program was in the order of 100's of millions of dollars in the 1970's."¹⁶ The CGIS was both a set of electronic maps and the "computer programs that allowed users to input, manipulate, analyse, and output those maps."¹⁷

By the late 1980s, the CGIS was no longer being used. Priorities changed, people retired, and institutional memory was fading. Numerous boxes of tapes and racks of documentation were left behind and only a few computers were capable of reading nine track tapes let alone run the programs. In 1995, an informal trans-organizational group of individuals from Statistics Canada, the *National Atlas of Canada*, Archives Canada and the private sector joined forces to restore it. Fortunately members of this salvage team had either formerly worked on the CGIS or had an interest in its preservation, which meant they had the requisite skills, knowledge, and more importantly they recognized the value of these data and knew that these were part of Canada's technological and scientific heritage that could be repurposed into other applications. On June 18, 1998, the agriculturally relevant portions of the CLI were handed over on one CD and it worked flawlessly on the analytical tools built in anticipation of the new format and eventually the CLI was distributed in the *GeoGratis*¹⁸ portal for free along with metadata and some text to help

¹² Schut, Peter. *Back from the Brink: the story of the remarkable resurrection of the Canada Land Inventory data*, (2000), available at: <http://web.archive.org/web/20011104160045/http://www.igs.net/~schut/cli.html> (accessed 20 August 2012).

¹³ Ahlgren, Dorothy and John McDonald. "The Archival Management of a Geographic Information System," *Archivaria* 13 (1981): 61.

¹⁴ Schut, *Back from the Brink*, (2000).

¹⁵ See the *GeoGratis* Canada Land Inventory description, available at <http://geogratis.cgdi.gc.ca/geogratis/en/collection/detail.do;jsessionid=155D0541488771F32F19DE8CDD39A37E?i d=CB6E057B-0D85-9B22-29DE-6351369A8B02> (accessed 20 August 2012).

¹⁶ Wilson, Cameron and Robert A. O'Neil. "GeoGratis: A Canadian Geospatial Data Infrastructure Component that Visualizes and Delivers Free Geospatial Data Sets," in *Proceedings of the International Cartographic Association Conference, Ottawa, 2009*, available from <http://icaci.org/publications/> (accessed 20 August 2012).

¹⁷ Schut, *Back from the Brink*, (2000).

¹⁸ *GeoGratis* disseminates geospatial data at no cost and without restrictions to the public. It is the first open data initiative in Canada and was the first to adopt an unrestricted user licence, it is available at <http://geogratis.cgdi.gc.ca/geogratis/en/index.html> (accessed 21 August 2012).

interpret the content. The CLI “rapidly became their most popular product.”¹⁹ Not all of the CLI data were saved and the cost of the effort described above was very substantial. The CLI has become the common basemap in *Atlas of Canada*²⁰ maps, layers are used to train satellite imagery software to recognize land cover patterns in remotely sensed images²¹ which are then used to inform the managing of forests and agricultural areas among many others resources.

The UK 1986 Domesday project, like the CGIS, was groundbreaking and a milestone in the history of multimedia computing. It was started by the BBC in 1984 to celebrate the 900-year anniversary of the original William of Normandy 1086 Norman Domesday Survey. The 1986 Domesday project, was conceived as an ‘electronic exhibition’ displaying British life much like the Great Exhibition of 1851.²² It was an interactive atlas that relied on crowdsourced VGI enlisting the help of 14,000 British schools and students. School microcomputers were mobilized to collect survey data, neighbourhood stories and pictures. These data “were combined with thousands of maps, still photographs and central statistical, written and visual information.”²³ The project was funded by the BBC and the European Commission.²⁴ A national disc contained interactive national statistics, photos, newspaper and magazine clippings, virtual reality tours, and movies. The community disc, or the ‘people’s database’ was compiled by nearly 1,000,000 people, contained four by three kilometer blocks of land with photos, data and text for 80 cities.²⁵ These multimedia data²⁶ were georeferenced to Ordnance Survey (OS) maps, airphotos and satellite images. The hardware consisted of a BBC microcomputer with floppy disk drives and a tracker-ball; Philips Laservision in Read Only Memory (ROM); a high-resolution colour monitor and a proprietary retrieval and analysis software.²⁷ It was an easy to use, information rich, interactive database driven multimedia second generation GIS. It also put state data into the public domain, stimulated much discussion about the emerging ‘information marketplace’ and the public versus the private sector’s rights to access public data.²⁸

Software and hardware, however, quickly became obsolete. In the late 1990s rescue and salvage work began and, just in time, as only a couple of fully operational systems were still operative. In 1999 a consortium under the name of CAMiLEON was formed to emulate the Domesday software into modern computers and between 2002-2003 they successfully did so. In 2001 a process to reverse engineer the

¹⁹ Schut, *Back from the Brink*, (2000).

²⁰ The online version of the *Atlas of Canada*, available at <http://atlas.nrcan.gc.ca/site/english/index.html> (accessed 20 August 2012).

²¹ To see how the CLI was used in the *Atlas of Canada*, see chapter 5 and the appendices of Lauriault, Tracey. *Data, Infrastructures and Geographical Imaginations*, Ph.D. diss., Carleton University, Ottawa, 2012.

²² See the discussion by the producers of the Domesday Project: Goddard, John and Peter Armstrong, “The 1986 Domesday Project,” *Transactions of the Institute of British Geographers, New Series* 11, no. 3 (1986): 290-295, available at <http://www.jstor.org/stable/621789> (accessed 20 August 2012).

²³ Darlington, Jeffrey, Andy Finney and Adrian Pearce. “Domesday Redux: The Rescue of the BBC Domesday Project Videodiscs,” *Ariadne* 36 (2003), available at <http://www.ariadne.ac.uk/issue36/tna/> (accessed 20 August 2012).

²⁴ For details about software, code, algorithms and hardware refer to the Darlington, Finney and Pearce (2003). For GIS innovation, geospatial data, functionality and development see David Rhind, Peter Armstrong and Stan Openshaw. “The Domesday Machine: A Nationwide Geographical Information System,” *The Geographical Journal* 154, no. 1 (1988): 56-68, available at <http://www.jstor.org/stable/633476> (accessed 20 August 2012).

²⁵ Rhind, Armstrong and Openshaw (1988), “The Domesday Machine: A Nationwide Geographical Information System.”

²⁶ 40,000 photographs, 22,000 maps, full Ordnance Survey Coverage at 1:50 000, see Goddard and Armstrong (1986).

²⁷ See Rhind, Armstrong and Openshaw (1988).

²⁸ Goddard and Armstrong (1986), pp. 294-295.

data began and by 2004, Adrian Pearce was able make them useable and accessible online.²⁹ And, in 2003 high quality copies of the original discs were made and the UK National Archives re-disseminated it.³⁰ Today the 1986 Domesday Project is available on the BBC *Domesday Reloaded*³¹ website while the content of the 1086 Domesday volumes is made available online with locations georeferenced with page citations using Google Maps in a project called the *Open Domesday*.³²

Both the 1960's Canada Land Inventory (CLI) and the 1986 Domesday Project were massive, expensive, national scale projects. They were technologically, scientifically and procedurally innovative marking a turning point in the history of computerization and science.³³ The CLI informed the creation of wildlife reserves, legislation to protect farmland, and were base maps in ecological frameworks which still in use today.³⁴ Both projects contributed to the geographical knowledge base of their respective nations and remain well-used, loved and important scientific records.³⁵ In both cases their data were inseparable from their software while the 1986 Domesday data could only be viewed with BBC proprietary hardware. Hardware, storage media, and software became obsolete and teams of dedicated people were able to restore these historical records and make them available to thousands of users today. These two examples illustrate the importance of considering long-term preservation at the point of creation by implementing good record keeping practices and to have a data management strategy in place as projects evolve.³⁶ Furthermore, they are arguments against the use of proprietary systems, and for the adoption of open specifications, standards and interoperability and for 'proactive archiving'.³⁷

²⁹ For the technical details on the rescue of the 1986 Domesday, see Darlington, Finney and Pearce (2003), "Domesday Redux."

³⁰ See the "Story of the Domesday Project," available on the *Domesday Reloaded* webpage at <http://www.bbc.co.uk/history/domesday/story> (accessed 20 August 2012).

³¹ The website is available at <http://www.bbc.co.uk/history/domesday/using-domesday> (accessed 20 August 2012).

³² *Open Domesday* is available at <http://domesdaymap.co.uk/> (accessed 20 August 2012).

³³ See R. J. Morris, "History and Computing: Expansion and Achievements," *Social Science Computer Review* 9, no. 2 (summer 1991): 215-230.

³⁴ The *Canada Encyclopedia* entry on the CLI provides describes of how it has contributed to the advancement of the Canadian economy and society, available at <http://www.thecanadianencyclopedia.com/articles/canada-land-inventory> (accessed 20 August 2012).

³⁵ There were 45,373 CLI map downloads in 2011-2012 and The Province of Ontario has a very popular CLI product available on this website <http://www.omafra.gov.on.ca/english/landuse/feed-in-tariffprogram.htm#3>, email communication received from Odette Trottier, Natural Resources Canada, Centre for topographic information on 28 August 2012.

³⁶ The InterPARES 2 Case Study on the *Cybercartographic Atlas of Antarctica* demonstrated that it is possible to design these aspects into systems at the point of creation, CS06 is available at http://www.interpares.org/ip2/ip2_case_studies.cfm?study=5. The IPY project mandated that all contributing scientists implement a data management strategy as part of funded projects see the Canada *IPY Data Management* at http://www.api-ipy.gc.ca/pg_IPYAPI_052-eng.html. Others have written about the technological issues of open source, standards and specifications, see Andrew Williamson, "Strategies for managing digital content formats," (2005) *Library Review* 54, no. 9: 508-513, and the Open Geospatial Consortium Data Preservation Working Group at <http://www.opengeospatial.org/projects/groups/preservdwg> (accessed 21 August 2012).

³⁷ In the case of cybercartographic atlases the use of open source, adherence to interoperable and open specifications, metadata as well as storing of multiple copies in geographically dispersed servers ensures ongoing backup future accessibility. See Peter Doorn and Heiko Tjalsma, "Introduction: archiving research data," *Archival Science* 7 (2007): 1-20, doi:10.1007/s10502-007-9054-6. Also, the IP2, CS06 *Cybercartographic Atlas of Antarctica* Case Study co-authored by Tracey P. Lauriault and Yvette Hackett (2005), available at http://www.interpares.org/ip2/ip2_case_studies.cfm?study=5 (accessed 18 August 2012).

3. Cybercartographic atlases as collective memory systems and community archives

While the CLI and the 1986 Domesday project exemplify the rescue and salvage efforts of legacy artefacts, there are countless examples of equally significant and technologically innovative digital maps and atlases currently being created that would be archived if a preservation infrastructure existed. The Geomatics and Cartographic Research Centre (GCRC) for instance has been actively engaged in the creation of cutting edge cybercartographic atlases for over a decade. Cybercartography “is the organization, presentation, analysis and communication of spatially referenced information on a wide variety of topics of interest and use to society in an interactive, dynamic, multimedia, multisensory and multidisciplinary format.”³⁸ Cybercartography also offers an unprecedented opportunity for fundamentally rethinking the way we design, produce, disseminate and use maps on the Internet, both theoretically and in practice.

The GCRC has been applying the concepts of cybercartography to map traditional knowledge in Canada’s north, traditional place names with indigenous heritage and educational institutions and geonarratives³⁹ surrounding treaty processes. These atlases embody the collective memories⁴⁰ of those who have contributed to their creation and have become a means to record the historical, geographical, cultural and scientific facts that have been transmitted orally for centuries. These atlases are the first official recordings of this aurally transmitted knowledge and elders and communities have authoritatively endorsed each record. The communities who have contributed to and authorized them regard these atlases as living archives.

The Inuit siku (sea ice) Atlas⁴¹ for example was developed to respond to Inuit elders’ and hunters’ expressions of interest: to share their knowledge with youth; see more Inuit knowledge and northern content in the northern education system and to share their knowledge more broadly with scientists and the general public. The SIKU Atlas was compiled and developed to reflect the knowledge, stories, maps, language, and lessons shared through years of interviews, focus groups, sea ice trips, and workshops with local sea ice experts in Cape Dorset, Igloolik, Pangnirtung, and Clyde River, Nunavut. Interactive atlas features are used to enable students to explore and learn about various sea ice topics, maps, Inuktitut terminology, community-specific information, and project background, including audio, video, pictures, text, and maps.

³⁸ See D. R. F. Taylor, Key Note Address titled “Maps and Mapping in the Information Era,” in *Proceedings 18th International Cartographic Conference*, vol. 1, ed. L. Ottomson Stockholm (Galve: Swedish Cartographic Society, 1997), 1-10; and *The Concept of Cybercartography*, Public Lecture University of Redlands and ESRI Staff. Redlands, (2003). CA, USA.

³⁹ See Sebastien Caquard, “Cartography I: Mapping Narrative Cartography,” *Progress in Human Geography*, November 7, 2011, doi:10.1177/0309132511423796.

⁴⁰ Anthea Josias describes a number of black South African collective memory projects capturing stories in text, audio and video; remaking district maps and communities documenting apartheid histories and the post-apartheid reconciliation processes. See “Toward an understanding of archives as a feature of collective memory,” *Archival Science*, vol. 11, (2011) pp. 95–112, doi:10.1007/s10502-011-9136-3. Elizabeth Nannelli discusses the Comissão de Acolhimento, Verdade e Reconciliação de Timor-Leste (Commission for Reception, Truth and Reconciliation, or CAVR) that recorded the testimonies of East Timorese living under Indonesian rule for 25 years between 1974 and 1999. See the “Records of the Commission for Reception, Truth, and Reconciliation in Timor-Leste,” *Archival Science* 9 (2009): 29-41, doi:10.1007/s10502-009-9103-4. Like cybercartographic atlases, these projects are making records outside of traditional archives and have become key references and the means to transfer knowledge by and for the communities that have created them. Only the CAVR is formally archived.

⁴¹ The Inuit siku (sea ice) Atlas (<http://sikuatlas.ca/>) was developed as part of an International Polar Year project called The Inuit Sea Ice Use and Occupancy Project (ISIUOP), through the collaboration of many northern, academic, government, and private industry contributors. It is a compilation of Inuit sea ice knowledge and use, as documented between 2004 – 2008; it is available at <http://sikuatlas.ca/index.html> (accessed 27 August 2012).



Figure 1. The Inuit Siku (sea ice) Atlas.

The Kitikmeot Heritage Society, Inuit Heritage Trust's Traditional Name Placing Project and the Gwich'in Cultural Society among others have approached the GCRC to help them map their place name databases with their collections of audio recordings of place names and video recordings of elders narrating the stories of those places. These place name atlases⁴² have become important knowledge transmission tools from elder to youth, are cultural geo-linguistic heritage preservation tools, have been incorporated as part of school curricula and are land occupancy records depicting the territorial extent of a community's land use and settlement. Naming places is part of the infrastructure of experience⁴³ and represents social relationships, kinship, historical events and shared cultural memories. These atlases are enabling local communities to replace their histories onto the map and others to see space from a different cultural lens.

The *Cybercartographic Atlas of the Lake Huron Treaty Relationship Process* (CALHTRP)⁴⁴ on the other hand is a retelling and re-mapping of the treaty process whereby historical archival records such as surveyor notebooks and diary entries, and numerous personal histories are geo-transcribed⁴⁵ and

⁴² See the *Kitikmeot Place Name Atlas* <http://www.kitikmeotheritage.ca/atlas.htm>, *The Arctic Bay Atlas* <http://arcticbayatlas.ca/index.html> (accessed 27 August 2012) and the *Gwich'in Goonanh'kak Goonwandak: The Places and Stories of the Gwich'in* (under development).

⁴³ See Paul Dourish and Genevieve Bell, "The Infrastructure of Experience and the Experience of Infrastructure: Meaning and Structure in Everyday Encounters with Space," *Environment and Planning B: Planning and Design* 34 (2005): 414 – 430.

⁴⁴ Available at <https://gcr.ca/confluence/display/GCRCWEB/The+Cybercartographic+Atlas+of+the+Lake+Huron+Treaty+Relationship+Process> (accessed 29 August 2012).

⁴⁵ Stephanie Pyne coined this term during the making of the *Cybercartographic Atlas of Indigenous Perspectives* in, "A "living Atlas" for Geospatial Storytelling: The Cybercartographic Atlas of Indigenous Perspectives and Knowledge of the Great Lakes Region," *Cartographica* 44, no. 2 (2009): 83-100.



Figure 2. Gwichin Place Name Atlas.

aggregated into a database. This database of stories is then mapped onto old and new maps along with georeferenced photographs, historical maps, signed treaties and the historical and archival records surrounding them. In addition, some historical maps were geo-rectified according to contemporary map projections⁴⁶ and layered with the oral history of events accounted by community elders or with the geo-transcribed stories from surveyor and explorer notes. These participatory and counter cartographies are enacting a post colonial mapping of Canada's treaty system. These atlases are therefore 'remapping' the official historical record by reflexively using western geospatial technologies, multimedia and methods in such a way so as not to recolonize.⁴⁷ Furthermore, they are also providing geonarratives to the making of colonial maps, making new records from old oral traditions, and by doing so put into question the 'authenticity' and 'completeness' of the traditional archival treaty record. Also, by geo-transcribing historical records, digitizing oral cultures, and re-purposing old maps in new ways, cybercartographic atlases are giving each of these records a 'secondary provenance',⁴⁸ even though the provenance of the

⁴⁶ Recently historical maps in the David Rumsey Map Collection have been geo-rectified to be viewed along in Google Map and Google Earth, available at <http://rumsey.geogara.com/gmaps.html> (accessed 29 August 2012).

⁴⁷ For a more detailed account of this process see Stephanie Pyne and D.R. Fraser Taylor, "Mapping Indigenous Perspectives in the Making of the Cybercartographic Atlas of the Lake Huron Treaty Relationships Process: A Performative Approach in a Reconciliation Context," *Cartographica* 47, no. 2 (2012): 92-104.

⁴⁸ Lori Podolsky Nordland in "The Concept of "Secondary Provenance": Re-interpreting Ac ko mok ki's Map as Evolving Text" discusses transmedia shifts of pre-Gutenberg archival records and how these digitized records gain new life and acquire new layers of meaning. Furthermore, in the case of the Ac ko mok ki' map, it was argued that it



Figure 3. Cybercartographic Atlas of the Lake Huron Treaty Relationship Process.

original record is provided, it becomes less important as the atlases allow for the re-interpretation of original records.

Aboriginal social memory or traditional knowledge is now being recognized in the courts for evidential purposes,⁴⁹ to countervail archival sources⁵⁰ and there is growing recognition that archival records such as treaties, maps and surveys were modernist legal devices of British and French colonialists. As Raymond Frogner noted, these records represented “sovereignty’s positivisation” and were legal fictions since they purported to have been created among equals,⁵¹ which was mostly not the case. The CALHTRP in particular, puts into question the ideas of treaties as dispositive documents as the transaction did not necessarily involve willful participants.⁵²

should be reinterpreted according to the Siksika world view and their cartographic conventions thus giving it a secondary provenance, *Archivaria* 58 (2004): 147-159.

⁴⁹ A famous Canadian example is the case of the Gitksan and Wet’suwet’en First Nations who used their oral histories, which capture their geo-histories in song, performances and stories in a land rights trial against the Province of British Columbia and the federal government. They also translated their traditional knowledge into a map. Furthermore, the plaintiffs, decided that to make their claim they “had to turn the legal system, its archives, precedents, and process against itself” as they recognized they were playing in a fixed game see p. 47 of Mathew Sparke, “A Map that Roared and an Original Atlas: Canada, Cartography and the Narration of Nation,” *Annals of the Association of American Geographers* 88, no. 3 (1998): 463-495.

⁵⁰ Raymond Frogner, in “Innocent Legal Fiction: Archival Convention and the North Saanich Treaty of 1852” demonstrated “that conventional archival interpretations identify the silences and discrepancies of the textual colonial record. But critics note that conventional archival method remains tied to its textual and sovereign paradigms. It does not address the often vague and uncertain relationship between the record and the manifold power structures, cultures, and traditions that surround the record’s formation and archival disposition,” *Archivaria* 70 (Fall 2010): 45.

⁵¹ Frogner, *Archivaria* 70 (Fall 2010): 47.

⁵² Oral histories surrounding the *North Saanich Treaty* are featured by Frogner (Fall 2010), pp. 81-86 and dispositive documents are discussed on p. 48.

In these three atlas examples Aboriginal elders and communities appraised the content of the maps; and also supplied, made and curated the records; and directed how these were to be cartographically rendered in collaboration with researchers and programmers at the GCRC. These atlases have become a community archive created outside the traditional archive with the use of ‘authoritative’ maps, methods and technologies, archival records, and by georeferencing traditional knowledge from ‘authoritative’ community sources onto new and old maps. In a sense they have put into question the authenticity of what ‘official’ archival documents about these spaces purport to be and have created a counter geo-narrative.

4. The Preservation of Geospatial Data

The Government of Canada has a number of ongoing geospatial data preservation strategies as seen in Table 1.⁵³ These thematic geospatial data archives are part of science-based departments and data are collected as part of the business function and reporting responsibilities of their custodial institutions. Also, these data are collected according to established scientific data methods that are deeply ingrained practices embedded into the tools from which data are sensed, how quality is assessed, and in how these data are organized, described, formatted, disseminated and used.⁵⁴ Of all the initiatives, only the Department of Fisheries and Oceans (DFO) archived data are part of a departmental science strategy. These archives are also collaborative endeavours which contain data accessioned in partnership with others, are collected as part of a cost-recovery arrangement with other institutions, are the holdings of partner institutions, are part of major collaborative research projects, are derived from sensors housed and managed in many institutions, or are derived from a myriad private and public sector satellite networks. All but the NSDB, which contains legacy datasets, include a combination of data from active and inactive sensors, which are continuously being updated, in many cases with near-real time data. They are therefore growing collections of data that have not necessarily been set aside for disposition purposes into a traditional archive and may remain part of ongoing business practices.

Many of the archives examined disseminate and allow for the visualization of their data using specialized software while none mention the preservation of software, hardware and associated specialized file formats. These initiatives may however not be archives in the traditional sense, because they are not “an agency or institution responsible for the preservation and communication of records selected for permanent preservation” or a “place where records selected for permanent preservation are kept.”⁵⁵ It is uncertain if the institutions within which the *Earth Observation Data Services (EODS)*, *The National Soil DataBase (NSDB)* or the *Integrated Science Data Management (ISDM) Wave Data Archive* housed respectively at NRCan, AAFC or DFO, have a data preservation policy in place together with the

⁵³ The following information forms part of a larger study commissioned by GeoConnections entitled *TA 2: Final Report: Geospatial Data Archiving and Preservation* as part of the Science & Technology Policy Research and Analysis Resource Team Prepared for: NRCan, GeoConnections Operational Framework Team by Hicklings, Arthur and Low (HAL), authored by Tracey P. Lauriault and Ed Kennedy, (March 2011).

⁵⁴ The IP2 Scientific Data Portal General Study discussed the specificities of disciplinary and sub-disciplinary scientific normative practices available at http://www.interpares.org/ip2/ip2_case_studies.cfm?study=34. Also see The Focus 2 – The Sciences section of Yvette Hackett, William Underwood, and Philip Eppard, “Part One: Case and General Studies in the Artistic, Scientific and Governmental Sectors Focus Task Force Report,” *InterPARES 2: Experiential, Interactive and Dynamic Records*, ed. Luciana Duranti and Randy Preston (2008), http://www.interpares.org/display_file.cfm?doc=ip2_book_part_1_focus_task_force.pdf (accessed 29 August 2012).

⁵⁵ Def. Archive: InterPARES 2 Terminology Dictionary, http://www.interpares.org/ip2/ip2_terminology_db.cfm.

necessary human and technological resources required to ensure the permanent preservation of these data and databases. These databases may simply be backed up. The cost recovery potential of the *EODS* data would warrant investment in a good storage, retrieval and backup system, while the *NSDB* could simply be a collection of agricultural data considered important by a group of dedicated soil scientists who have the skill and will to be their custodians. The *ISDM Wave Data Archive*, may be the closest to being a ‘traditional archive’ since DFO has a *Management Policy for Scientific Data*⁵⁶ that includes a mandate, model, infrastructure, and human resources dedicated to the archiving and preservation of their data resources. It is presumed that the DFO *IOS/OSD Data Archive* would fall under the same policy.

In terms of a preservation strategy,⁵⁷ these initiatives do share characteristic elements of a preservation repository, such as:

- The means to access the data and metadata;
- Reference and context information;
- Provenance information;
- Licensing and terms of use;
- File format information; and
- In some cases are created within a data management policy (e.g., DFO).

Most geospatial data portals that store and manage the data they disseminate (e.g., *GeoBase* and *GeoGratis*) have similar characteristics. It is harder to assess file transfer mechanisms, preservation strategies such as data migration, emulation, etc., file name conventions, unique identifiers, adherence to standard vocabularies, backup schedules and the technology used from their websites.

While imperfect in terms of a traditional archival perspective, these are nonetheless starting points to build a more comprehensive geospatial data and mapping preservation strategy. These however only represent but a very small fraction of the geospatial data produced at NRCan and the rest of the Government of Canada.⁵⁸ They are also examples of proactive archiving practices—some of which are inherent in geospatial data portals and of databases considered being part of a records management process.

5. Geospatial Data Management Models

The DFO Integrated Science Data Management⁵⁹ and the International Polar Year (IPY) Data and Information Service have developed data management models, which take into consideration preservation.

⁵⁶ DFO Management Policy for Scientific Data available, <http://www.dfo-mpo.gc.ca/science/data-donnees/policy-politique-eng.htm> (accessed 22 August 2012).

⁵⁷ Defined as “a coherent set of objectives and methods for protecting and maintaining (i.e., safeguarding authenticity and ensuring accessibility of) digital components and related information of acquired records over time, and for reproducing the related authentic records and/or archival aggregations,” in Def. Records Preservation Strategy: InterPARES 2 Terminology Database available at http://www.interpares.org/ip2/ip2_terminology_db.cfm.

⁵⁸ For a more detailed analysis of these initiative according to the 10 principles of digital preservation repositories developed by the Center for Research Libraries in Chicago see the HAL (2011) *TA 2: Final Report: Geospatial Data Archiving and Preservation*.

⁵⁹ DFO Integrated Science Data Management (ISDM), <http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/index-eng.html> (accessed 21 August 2012).

The *ISDM Wave Data Archive* (See Table 1) integrates the preservation and archiving of data from multiple organizations into an operational information management environment. The active field program of wave acquisition started in 1971, eventually discontinued in 1996, the process of submitting delayed-mode wave data has continued. In addition, ISDM acquires daily wave data from buoys operated by the Meteorological Service of Canada (MSC) at EC and manages scientific data from a wide range of sources and contains over 10 million hourly sea state and swell measurements from some 500 locations in Canada's lakes and surrounding oceans. While these data are being used in real-time to produce weather and sea state forecasts, its archiving program ensures that the information can be used in a variety of applications requiring data over long timeframes, such as hindcast models of wave climatology used in ocean maritime navigation, engineering and climate change studies.

Table 1. Government of Canada geospatial data preservation initiatives

IOS/OSD Data Archive , Department of Fisheries and Oceans (DFO)	National Climate Data and Information Archive , Environment Canada (EC)	Integrated Science Data Management (ISDM) Wave Data Archive , DFO
Lithoprobe Data Archive , NRCan, Geological Survey of Canada (GSC)	The Canadian Ice Service Archive (CISA) , EC, Canadian Ice Service	National WaveForm Archive (NWFA) , NRCan, Earthquakes Canada
The National Soil DataBase (NSDB) , Agriculture and Agri- Food Canada (AAFC); Canadian Soil Information Service (CANSis)	Earth Observation Data Services (EODS) , Natural Resources Canada (NRCan), Canada Centre for Remote Sensing (CCRS)	National Water Data Archive , EC, Water Survey of Canada (WSC) w/Archived Sediment Data
Geomagnetism Summary Plots: Archives , NRCan, GSC	System of Agents for Forest Observation Research with Advanced Hierarchies (SAFORAH) , Canadian Forest Service (CFS), University of Victoria and other academic and government partners.	

The archiving and preservation commitment extends across all kinds of scientific data for which DFO is responsible, as evidenced by the department's *Management Policy for Scientific Data*,⁶⁰ which came into effect in June 2001. This IM Policy includes several references to the requirement for data archiving and preservation and it adheres to the following principles:

1. DFO scientific data sets... are irreplaceable, and must be protected and managed to ensure long-term availability;
2. ... it is essential that DFO Science/Oceans maintain responsibility for their quality control, management, archiving and dissemination; and
3. ... all scientific data collected by the DFO must be migrated to a 'managed' archive immediately after the data have been processed.

⁶⁰ DFO, *Management Policy for Scientific Data*, (2001), <http://www.dfo-mpo.gc.ca/science/data-donnees/policy-politique-eng.htm> (accessed 21 August 2012).

In terms of archiving, all DFO scientific data must be managed as part of an integrated system accessible through regional, zonal and national data centres and the responsibilities of the integrated system of data centres will be to:

- Ensure long-term accessibility and documentation in the event of organizational changes, retirements, etc.;
- Protect data against loss resulting from error, accident, technological change, degradation of media, etc.

When data are submitted, the DFO stresses the importance of ensuring that data are quickly migrated into a 'managed' environment where they are properly backed up and secured from accidental or circumstantial loss, and where the supporting metadata are integrated with the data to preserve their long-term usefulness. Finally the DFO includes a data rescue program to locate and preserve scientific data that are of value to departmental programs and to identify data at risk.

The International Polar Year (IPY) project, on the other hand, was a large three-year Arctic and Antarctic scientific program, organized through the International Council for Science (ICSU) and the World Meteorological Organization (WMO). The IPY Data and Information Service (IPYDIS)⁶¹ was a global partnership of data centers, archives, and networks working to ensure the proper stewardship of IPY and related data. The National Snow and Ice Data Center (NSIDC) at the University of Colorado is a coordination office for the IPYDIS and it ensures the long-term preservation and access to IPY data. The policy⁶² states that:

...it is essential to ensure long-term preservation and sustained access to IPY data. All IPY data must be archived in their simplest, useful form and be accompanied by a complete metadata description. An IPY Data and Information Service (IPYDIS) should help projects identify appropriate long-term archives and data centers, but it is the responsibility of individual IPY projects to make arrangements with long-term archives to ensure the preservation of their data. It must be recognized that data preservation and access should not be afterthoughts and need to be considered while data collection plans are developed.

The IPY exemplifies a distributed or "virtual" archive of scientific data and proactive archiving. The Service's Web site guides digital scientific data users to a number of portals and centers that are currently providing access to IPY and related data and all data registries and repositories are required to adhere to the IPY Metadata Profile requirements. This profile was based on the Global Change Master Directory (GCMD) Directory Interchange Format, and is compliant with and has been mapped to recognized geospatial metadata standards.⁶³

⁶¹ IPY, *International Polar Year Data and Information Service (IPYDIS)*, Ironically the IPYDIS site is not longer supported but copies of key documents are available from Mark A. Parsons at parsonsm@nsidc.org.

⁶² IPY, (2008), *International Polar Year 2007-2008 Data Policy* is available at http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf (accessed 21 August 2012).

⁶³ The standards are: Global Change Master Directory (GCMD) Directory Interchange Format (DIF); the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) (FGDC-STD-001-1998) and Remote Sensing Extensions (RSE) (FGDC-STD-012-2002); and THREDDS Dataset Inventory Catalog Specification Version 1.0. Other international standards are recommended such as the Open Geospatial Consortium (OGC) standards, and the ISO 19115:2003 metadata standard. See IPY Data and Information Management Service, *IPY Metadata Profile version 1.0*, available from Mark A. Parsons parsonsm@nsidc.org.

The DFO and IPY policies, practices and initiatives, like the archiving examples just discussed above, are good places to start when planning a strategy to preserve geospatial data and maps.

6. Canadian digital data and information consultations, studies, reports and initiatives

The above-mentioned geospatial data archives and data management models occurred in spite of the paucity of tangible outcomes from the 12 consultations, studies, reports and initiatives as seen in figure 4 below.

The *Research Data Strategy*⁶⁴ is one of the tangible outcomes of the 2002 *NCSARD* report process that will be discussed shortly. It is a collaborative effort between government and academia aiming to address the challenges and issues surrounding access to and preservation of data created by Canadian researchers. It is a multidisciplinary group of universities, institutes, libraries, granting agencies, and individual researchers, including several officials involved with the creation, management, dissemination and/or funding of geospatial data,⁶⁵ who recognize the need to address data management issues. While the focus is on research data, the ideas are transferable to the preservation and management of geospatial data. The approach and principles to data stewardship recommended by *Research Data Canada* have been echoed in many other initiatives discussed here, namely that data stewardship and management needs to be framed in a life cycle model at all stages of the data production process and making data

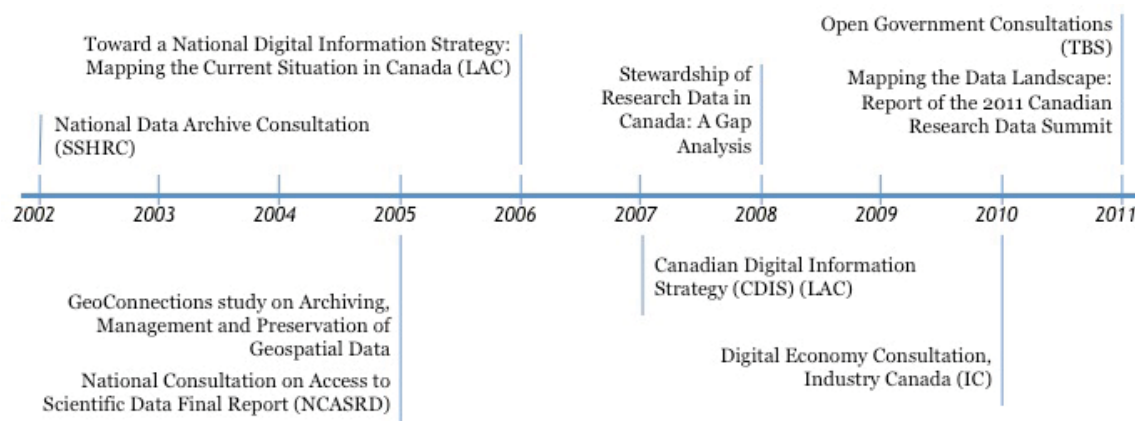


Figure 4. Canadian Digital Data and Information Consultations, Studies, Reports and Initiatives.

preservation a formal institutional responsibility.⁶⁶

Two international groups hold potential for future guidance in geospatial data archiving and preservation. The Committee on Data for Science and Technology (*CODATA*) *Working Group on*

⁶⁴ Available at the Research Data Strategy home page, <http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/> (accessed 27 August 2012).

⁶⁵ John Broome, the Chair CNC/CODATA and Head, Data Management Policy and Strategy, Earth Sciences Sector, Natural Resources Canada; Scott Tomlinson, Data Management Coordinator, International Polar Year Federal Program and Chuck Humphrey, Data Library Coordinator, Libraries, University of Alberta.

⁶⁶ Research Data Canada, <http://ncasrd-cnads.scitech.gc.ca/eng/about/principles-data-stewardship.html>.

*Archiving Scientific Data*⁶⁷ has been holding symposia and workshops on the topic and the *Canadian National Committee for CODATA*⁶⁸ has been active in documenting and reporting scientific data activities. Members of these include geospatial data experts. The Open Geospatial Consortium (OGC) also has a *Data Preservation WG*,⁶⁹ which aims “to address technical and institutional challenges posed by data preservation, to interface with other OGC working groups that address technical areas that are affected by the data preservation problem, and to engage in outreach and communication with the preservation and archival information community.” Reports from this WG are not yet available.

In terms of consultation, only the 2011 TBS *Open Government Consultations*⁷⁰ yielded a report that informed an actual plan⁷¹ that was released at the Open Government Partnership in Brazil⁷² in the spring of 2012. The plan includes better record keeping and the lifting of restrictions to archived material at LAC but it does not include the preservation of data in the TBS *Open Data Pilot Project* portal nor any kind of preservation strategy for maps and data produced by government more broadly. The TBS Consultation report and the Plan also make no reference to preservation and archiving, LAC was not one of the ‘thought leaders’ to make a public submission and the complete list of individuals and organizations who submitted has not been made public.⁷³ Furthermore, the Advisory Panel on Open Government⁷⁴ does not include a representative from LAC, libraries, archives or science although the President and Founder of ESRI Canada is a member. Interestingly, the most popularly requested datasets in the TBS Consultation were geospatial and geospatial datasets predominate in terms of number of datasets in the *Open Data Pilot Project* portal.⁷⁵

The *Mapping the Data Landscape: Report of the 2011 Canadian Research Data Summit*⁷⁶ was the result of National Research Council (NRC) of Canada 2011 Canadian Research Data Summit held in September of 2011 and summit partners included Canadian IPY members.⁷⁷ The report is currently being circulated to discuss the *Research Data Strategy*.

The 2010 IC *Digital Economy Consultation*⁷⁸ did not result in any consolidated report or actions. The IC *Consultation Paper*⁷⁹ circulated to frame the consultation, however, did mention improving access

⁶⁷ *CODATA Preservation of and Access to Scientific and Technical Data in Developing Countries*, 2006, Home Page, <http://www.codata.org/taskgroups/TGpreservation/> (accessed March 2011).

⁶⁸ *Canadian National Committee for CODATA*, <http://dac.cisti.nrc.ca/> (accessed March 2011).

⁶⁹ *OGC, Data Preservation WG*, <http://www.opengeospatial.org/projects/groups/preservwg>) (accessed March 2011).

⁷⁰ Treasury Board Secretariat of Canada, December 6, 2011 to January 16, 2012, *Open Government Consultation Report* available at <http://www.open.gc.ca/consult/wwh-cr-eng.asp> (accessed 1 July 2012).

⁷¹ TBS Open Government Plan, released April 12, 2012, <http://www.tbs-sct.gc.ca/media/nr-cp/2012/0412-eng.asp#back1> (accessed 26 June 2011).

⁷² Available at <http://www.opengovpartnership.org/> (accessed 29 August 2012).

⁷³ Lauriault included preservation in her submission posted on the datalibre.ca blog on January 12, 2012, <http://datalibre.ca/2012/01/16/tbs-open-data-and-open-government-consultation-response/> (accessed 26 June 2011).

⁷⁴ Open Government Advisory Panel, <http://www.open.gc.ca/open-ouvert/bio-bio-eng.asp>.

⁷⁵ When the *Open Data Pilot Project* portal was launched, IASSIST reported that there were “782 general datasets (e.g., not geospatial) and over 260,000 geospatial datasets” available at <http://www.iassistdata.org/resources/canadas-open-data-pilot-project> (accessed 26 June 2011).

⁷⁶ The report is available at rds-sdr.cisti-icist.nrc-cnrc.gc.ca/docs/data_summit-sommet_donnees/Data_Summit_Report.pdf (accessed 27 August 2012).

⁷⁷ See the report background material available at http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/events/data_summit_2011/index.html (accessed 28 August 2012).

⁷⁸ Industry Canada, (2010), *Digital Economy Consultation*, available at http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/h_00491.html (accessed 27 August 2012).

to research data, data discovery and open access publishing but did not explicitly mention the geomatics sector or geospatial data. A cursory examination of submissions and the idea forum⁸⁰ called for the creation of a digital data and information archive and a network of trusted digital repositories.⁸¹ The summary of key findings⁸² however failed to mention these calls.

The *Stewardship of Research Data in Canada: A Gap Analysis*⁸³ was conducted by members of *Research Data Canada*. The life-cycle model was used as the framework to analyse the current state of research data stewardship in Canada, which includes: data production, dissemination, long-term preservation and data discovery and repurposing. The *Gap Analysis* reviewed: policies; funding; roles and responsibilities; repositories; standards; skill and training; rewards and recognition systems; R&D; access; and preservation. Not surprisingly, the analysis concluded that most research data created in Canada are greatly underutilized and are at a high risk of being lost. Geospatial and research data share similar characteristics, namely: they are complex, are in multiple data formats, use non standardized rendering techniques and specialized open and/or proprietary software, are disseminated in portals, or are shared and visualized in distributed systems, datasets can be very large and data can be modeled. Geospatial initiatives, unlike research data do have uniform metadata and data quality standards are normalized.

The 2007 LAC *Canadian Digital Information Strategy (CDIS)*⁸⁴ report reflects the broad consensus reached at a *National Summit*⁸⁵ held in 2006 which included contributions from more than 200 specialists some of whom are involved in geospatial data creation, use, dissemination and management. There are numerous suggested actions listed, none of which are specific to geospatial data, although GeoConnections, GeoGratis, the Canada Ice Service Data Archive and GeoBase are lauded for some of their initiatives. Environment Canada, Statistics Canada, DFO and NRCan are mentioned as institutions that produce and hold significant collections of government data. Some of the preservation assumptions upon which the *CDIS* was premised mirror contemporary geospatial data productions practices and issues. For example: technological change is constant; a distributed, interoperable, standards based, open and

⁷⁹ Industry Canada, 2010, *Improving Canada's Digital Advantage Strategies for Sustainable Prosperity: Consultation Paper on a Digital Economy Strategy for Canada*, accessed June 26, 2012

http://publications.gc.ca/collections/collection_2010/ic/Iu4-144-2010-eng.pdf.

⁸⁰ The most popularly voted idea was for access to and the preservation of government data assets, see

<http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/00172.html> (accessed 29 August 2012).

⁸¹ See the following submissions by Canadian Association of Research Libraries

http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/00349_4.html, universities

http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/00357_4.html,

http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/00402_4.html, two individuals

http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/h_00491.html,

http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/00284_3.html, institutes, federations, associations and council

http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/00398_4.html,

http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/00418_4.html,

http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/h_00491.html, and

http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/00440_4.html, (accessed 21 August 2012).

⁸² Summary of key findings is available at <http://www.ic.gc.ca/eic/site/sittgateway-portailstit.nsf/eng/00055.html> (accessed 28 August 2012).

⁸³ Research Data Canada, (2008), *Stewardship of Research Data in Canada: A Gap Analysis*, available by contacting Bronwen.Woods@nrc-cnrc.gc.ca.

⁸⁴ Library and Archives Canada (2007), *Canadian Digital Information Strategy*, available at <http://www.lac-bac.gc.ca/obj/012033/f2/012033-1000-e.pdf> (accessed February 2011).

⁸⁵ Library and Archives Canada (2006), *Toward a Canadian Digital Information Strategy: National Summit*, accessed February 2011 <http://www.collectionscanada.gc.ca/obj/012033/f2/012033-611-e.pdf>.

accessible strategy is key; stakeholders and communities of practice input are essential. The proposed actions could be incorporated into existing Government of Canada data portals.

The LAC 2006 *Toward a National Digital Information Strategy: Mapping the Current Situation in Canada*⁸⁶ states that “the stewardship of digital information produced in Canada is disparate and uncoordinated” and “the volume, diversity and complexity of digital information is growing exponentially” while the “technologies, standards and practices that will better ensure the ongoing accessibility and integrity of digital information are not yet consistently applied.” The report includes a brief description of the GeoConnections program and lists a number of other geospatial data initiatives at different levels of government in Canada in its Annexes. The report does not provide specific recommendations on how to manage these resources but it does suggest that the:

...preservation of digital information requires a management response, not just a technological response. The response must be active and sustained through time and it must address the preservation of digital information from both the strategic and tactical perspectives and within the context of the management frameworks that govern the businesses of organizations themselves. It must also be comprehensive and address all facets of the required preservation infrastructure: the policies that assign accountability, the standards, practices, procedures and technologies that enable the implementation of preservation strategies, and the people that make it all happen. These issues present a tremendous challenge and must be addressed in an inclusive and collaborative manner.⁸⁷

MacDonald and Shearer also provide a number of next steps toward the creation of a national digital information strategy, which resemble the collaborative strategies used to create the CGDI (e.g., multisectoral, departmental and multidisciplinary collaboration, etc.).

In 2005 GeoConnections conducted a study on *Archiving, Management and Preservation of Geospatial Data*⁸⁸ which provided a well rounded analysis of geospatial data preservation issues, such as: technological obsolescence; formats; storage technologies; temporal management; and metadata.⁸⁹ The study also provided a list of technological preservation solutions with their associated advantages and disadvantages and recommended “as the first step in ensuring the long-term preservation and retention of valuable resources, data producers must adopt an information life cycle management approach, which will ensure that their data will be managed proactively from creation to disposition.”⁹⁰ The recommendations are somewhat outdated in light of institutional changes at NRCan and of the new TBS Directives and Guidelines, irrespective, the study provides the following useful recommendations:

⁸⁶ MacDonald, J. and K. Shearer (2006), *Toward a National Digital Information Strategy: Mapping the Current Situation in Canada* (Ottawa: Library and Archives Canada), accessed February 2011 Summary: <http://www.lac-bac.gc.ca/digital-initiatives/012018-3200-e.html> and full Final Report: <http://www.collectionscanada.gc.ca/obj/012018/f2/012018-3200-e.pdf> (accessed 29 August 2012).

⁸⁷ See MacDonald and Shearer (2006), p. 50.

⁸⁸ GeoConnections (2005). *Archiving, Management and Preservation of Geospatial Data: Summary Report and Recommendations*, produced by David L. Brown, Grace Welch and Christine Cullingworth as an initiative of the Policy Advisory Node.

⁸⁹ The approach adopts the framework created by Bleakly, Denise R. (2002). *Long-Term Spatial Data Preservation and Archiving: What are the Issues?* Sand Report, SAND 2002-0107. Albuquerque, New Mexico: Sandia National Laboratories available at <http://prod.sandia.gov/techlib/access-control.cgi/2002/020107.pdf> (accessed 28 August 2012).

⁹⁰ GeoConnections (2005), *Archiving, Management and Preservation of Geospatial Data*, Page i.

- Organizations should define and implement policies and practices for the creation, use, retention, dissemination, preservation, and disposition of geospatial data. Building a business case for the creation of core geospatial data products is suggested.
- Organizations must establish authoritative responsibility centers that empower individuals with the ability to define and apply the information management principles required to ensure the integrity of an organization's geospatial data holdings. A custodianship model is an option, which provide a means to facilitate data management on the behalf of creators and users, and provide continuity in the delivery of a geospatial data infrastructure.
- Geospatial data preservation issues fall within the realm of a national information policy, and a national data management strategy. Working in partnership with the library and archives communities, government data producers need to standardize and adopt organizational policies and practices to govern the creation, use, retention, dissemination, preservation, and disposition of geospatial data to ensure its authenticity and integrity for as long as it is required for legislation, departmental statutes and other laws and policies. Key geospatial advisory boards, councils, and standards organizations should work with LAC to develop a series of geospatial data policies and practice.
- The Canadian Council on Geomatics (CCOG) should create a task force and invite interested stakeholders from the academic community, private sector and other federal and provincial/territorial agencies that can collaborate to develop priority policy areas identified above.

The 2005 *National Consultation on Access to Scientific Data Final Report (NCASRD)*,⁹¹ developed in partnership with the NRC, the Canada Foundation for Innovation (CFI), Canadian Institutes of Health Research and Natural Sciences and Engineering Research Council (NSERC), expressed the concern that: “no national data preservation organization exists, nor does Canada have any national data access strategy or policies. NCASRD participants expressed considerable concern about the loss of data, both as national assets and definitive longitudinal baselines for the measurement of changes over time.”⁹² The *NCSARD* report provides a comprehensive list of recommendations that include organizing a Data Force and the creation of a Data Canada secretariat.

The 2002 *SSHRC National Data Archive Consultation*⁹³ is *NSCARD*'s predecessor, primarily discusses data in the social sciences and humanities and the preservation of data created in the course of state funded research projects. The *Consultation* identified important institutions, infrastructures, management frameworks and data creators and calls for the creation of a national research data archive.

To date not much has happened in terms of implementation. Currently, LAC does not have a digital geospatial data archive and there is no national network of trusted digital repositories with the explicit mandate to ingest geospatial data, research, government or private sector created data. In light of recent budget cuts at LAC, proactive archiving and records management might be the best interim solution.

⁹¹ F. D. Strong and P. B. Leach (2005). *The Final Report of the National Consultation on Access to Scientific Data*, Ottawa: Government of Canada, it is available by contacting Bronwen.Woods@nrc-cnrc.gc.ca.

⁹² Strong, and Leach (2005), *The Final Report of the NCASRD*, page 2.

⁹³ Social Science and Humanities Research Council (2002), *Final Report: National Consultation on Research Data Archiving, Building Infrastructure for Access to and Preservation of Research Data*, http://www.sshrc-crsh.gc.ca/about-au_sujet/publications/da_finalreport_e.pdf (accessed 29 August 2012).

7. Legislation, Directives and Policies

It is beyond the scope of this paper to discuss all applicable legislation, directives and policies. Tables 2 and 3 list the overarching legislation, regulation, directives and policies which apply to geospatial data and Table 4 includes NRCan legislation that explicitly mention geospatial data, maps, note books, surveys or software that are to be preserved or are considered to be official government records. For a more detailed analyses of these please refer to the 2011 *TA 2: Final Report: Geospatial Data Archiving and Preservation*.

The Library and Archives Act (Table 2) is specific to LAC, while the others direct how Government of Canada information is to be managed and disseminated irrespective of the creating institution. These acts and regulations govern the preservation, management and dissemination of geospatial data, databases and maps, which, irrespective of their form are government information. These acts and regulations provide the context within which TBS and departmental guidelines, directives, and policies related to the management of government information are created.

Table 2. Overarching legislation and regulation that apply to geospatial data.

The Library and Archives of Canada Act (2004, c. 11)	Personal Information Protection and Electronic Documents Act (2000, c. 5)
<i>Copyright Act</i> (R.S., 1985, c. C-42)	Canada Evidence Act (R.S., 1985, c. C-5)
Access to Information Act (R.S., 1985, c. A-1)	Legal Deposit of Publications Regulations (SOR/2006-337)
<i>Privacy Act</i> (R.S., 1985, c. P-21)	Privacy Regulations (SOR/83-508).

Table 3 lists information management policies, directives, guidelines and standards at the federal level in Canada related to data archiving preservation in general and some refer to geospatial information specifically. The primary responsibility for such policies lies with the TBS. The *Standard on Geospatial Data* is the only document that addresses geospatial information domain. Its objective is “to support stewardship and interoperability of information by ensuring that departments access, use and share geospatial data efficiently and effectively to support program and service delivery.” It does not explicitly reference digital information archiving and preservation, although there is a linkage between it and LAC’s *File Format Guidelines for Preservation and Long-term Access*,⁹⁴ which recommends geospatial metadata standards for the preservation of and long-term access to digital geospatial information held by government organizations.

⁹⁴ Library and Archives Canada, *File Format Guidelines for Preservation and Long-term Access*, available at <http://www.collectionscanada.gc.ca/digital-initiatives/012018-2200-e.html> accessed March 2011

Table 3. Information policies, guidelines and standards.

TBS Directive on Recordkeeping ⁹⁵	Policy on Management of Information Technology ⁹⁶
TBS Directive on Information Management Roles and Responsibilities ⁹⁷	TBS Multi-Institutional Disposition Authority (MIDA) ⁹⁸
TBS Standard for Electronic Documents and Records Management Solutions (EDRMS) ⁹⁹	LAC Guidelines: Local Digital Format Registry (LDFR) ¹⁰⁰
TBS Policy on Information Management ¹⁰¹	TBS Standard on Geospatial Data ¹⁰²
TBS Policy Framework for Information and Technology ¹⁰³	TBS Standard on Metadata ¹⁰⁴
Forthcoming Methodology associated with TBS Directive on Recordkeeping, which NRCan Records Management Group contributed.	

Table 4 lists acts and regulations that specifically address geospatial information and are under the responsibility of the Minister of Natural Resources.¹⁰⁵ The *Remote Sensing Space Systems Act* and the *Remote Sensing Space Systems Regulations* were added since they impact how the Canada Centre for Remote Sensing (CCRS) manages its geospatial data sets, while the *Charts and Nautical Publications Regulations* addresses marine geospatial information. Acts and regulation mention raw data, how these are to be processed, the systems used to view and process the data and the use of cryptography. There are also mentions of registrars and catalogs. In some there are specific references to archiving, deposition, data protection plans, the number of years data are to be kept and data curation. Finally, because these geospatial datasets are critical to the management of Canada's boundaries, the health and safety of Canadians, Canada's economy and the attestation of lands, these government records can all be called into evidence to resolve international and national disputes, or any other judicial proceeding. Clearly, the geospatial datasets mentioned here must be well managed once they are created and should be preserved accordingly. The Government of Canada produces geospatial data, databases and maps in many other

⁹⁵ See <http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?section=text&id=16552>

⁹⁶ See <http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?section=text&id=12755>

⁹⁷ See <http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?section=text&id=12754>

⁹⁸ See <http://www.collectionscanada.gc.ca/government/disposition/007007-1062-e.html> and <http://www.collectionscanada.gc.ca/government/disposition/index-e.html>

⁹⁹ See <http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=18910§ion=text>

¹⁰⁰ See <http://www.collectionscanada.gc.ca/digital-initiatives/012018-2200-e.html>

¹⁰¹ See <http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?section=text&id=12742>

¹⁰² See Treasury Board of Canada Secretariat, (2009). *Standard on Geospatial Data*, <http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?evttou=X&id=16553§ion=text> (accessed 29 August 2012)

¹⁰³ See <http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=12452§ion=text>

¹⁰⁴ Treasury Board of Canada Secretariat, (2010), *Standard on Metadata*, <http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=18909§ion=text> (accessed 29 August 2011).

¹⁰⁵ Natural Resources Canada, (modified 2010), *List of acts for which the Minister of Natural Resources is responsible*, <http://www.nrcan-rncan.gc.ca/com/resoress/actacte-eng.php>, (accessed 29 August 2012).

departments and agencies and each of these would have to adhere to acts and regulations for which they are responsible along with those in Tables 2 and 3.

Table 4. Acts and regulation governing geospatial data.

Department of Natural Resources Act (1994, c. 41)	Energy Monitoring Act (R.S., 1985, c. E-8)
Resources and Technical Surveys Act (R.S., 1985, c. R-7)	National Energy Board Act (R.S., 1985, c. N-7) and National Energy Board Electricity Regulations (SOR/97-130)
Canada Lands Surveys Act (R.S., 1985, c. L-6)	Northern Pipeline Act (R.S., 1985, c. N-26)
Forestry Act (R.S., 1985, c. F-30) and Regulations Respecting the Report on the State of Canada's Forests (SOR/95-479)	Nuclear Energy Act (R.S., 1985, c. A-16)
Canada Oil and Gas Operations Act (R.S., 1985, c. O-7) and Canada Oil and Gas Geophysical Operations Regulations (SOR/96-117)	Nuclear Fuel Waste Act (2002, c. 23)
Canada Petroleum Resources Act (1985, c. 36 (2 nd Supp.))	Nuclear Safety and Control Act (1997, c. 9)
Canada-Newfoundland Atlantic Accord Implementation Act (1987, c. 3)	Uranium Mines and Mills Regulations (SOR/2000-206)
Canada-Nova Scotia Offshore Petroleum Resources Accord Implementation Act (1988, c. 28)	Remote Sensing Space Systems Act (2005, c. 45) and Remote Sensing Space Systems Regulations (SOR/2007-66)
Energy Efficiency Act (1992, c. 36)	Charts and Nautical Publications Regulations (SOR/95-149)

The legislation reviewed in the HAL 2011 Report, demonstrated the diversity of types of geospatial data and maps, their forms, formats and associated software and systems, how they are managed, the context within which they are created, and how they are considered authentic. Also, many geospatial datasets are very highly regulated. This analysis, combined with an examination of data IP2 science data portals below, reaffirms the heterogeneous nature of geospatial data and maps and the importance of creators to be involved in the process of preserving geospatial data assets.

8. LAC Guidelines pertaining to Geospatial Data and Maps

Few LAC guidelines apply to cartographic material. The 2011 *Managing Cartographic, Architectural and Engineering Records in the Government of Canada*¹⁰⁶ mentions that geomatics records include systems, discs, CD-ROMs and other cartographic material in electronic formats. However, it refers to the

¹⁰⁶ Library and Archives Canada (LAC), *Managing Cartographic, Architectural and Engineering Records in the Government of Canada*, Ottawa: Government of Canada, <http://www.collectionscanada.gc.ca/government/002/007002-2050-e.html> (accessed 29 August 2012).

2001 Canadian Committee on Archival Description (CCAD) *Rules for Archival Description Chapter 5*,¹⁰⁷ which primarily address paper maps although general issues pertaining to digital databases and programs are covered in the 2003 *Chapter 9: Records in Electronic Form*.¹⁰⁸ The LAC Local Digital Format Registry (LDFR) *File Format Guidelines for Preservation and Long-term Access Version 1.0*¹⁰⁹ includes *Geospatial* guidelines and recommends TC 211 ISO 19115 Geographic Information—Metadata and input from digital geospatial data record creators and managers is solicited. These guidelines are, however, inadequate for the kind of complex geospatial artefacts discussed in InterPARES 2 Project (IP2) or for cybercartographic atlases.

9. Research on the preservation of geospatial data in Canada

The following discusses empirical examples from a selection of InterPARES 2 Project¹¹⁰ Case Studies¹¹¹ (Table 5) that study geospatial data, and part of the IP2 General Study 10 (GS10),¹¹² which examined geospatial data portals (Table 6).

The IP2 Case Studies in Table 5 below include observational data in a variety of forms, while five also included computational data or models. All but the Cybercartographic Atlas¹¹³ render or store their data in a proprietary system, and data are stored in a variety of databases or in a searchable data portal while Engineering project data¹¹⁴ (CS19), the Land Registry¹¹⁵ (CS18), and some of the NASA data (CS08) are inseparable from the systems within which they have been created and/or stored. As discussed earlier when examining the CLI and the 1986 Domesday project, these findings are somewhat discouraging.

The IP2 Case Study questions regarding authenticity¹¹⁶ revealed that the prevalent concept of reliability is closely tied to reproducibility and accuracy in the physical sciences. In many cases, reliability is associated with faith in the technological systems in place or security measures related to access to the system. In terms of authenticity, it was revealed that Case Study respondents¹¹⁷ do not think

¹⁰⁷ Canadian Committee on Archival Description, (2001), *Rules for Archival Description Chapter 5 Cartographic Materials*, http://www.cdncouncilarchives.ca/rad_ch5.pdf (accessed February 2011).

¹⁰⁸ Canadian Committee on Archival Description, (2003), *Rules for Archival Description Chapter 9 Records in Electronic Form*, (accessed February 2011). http://www.cdncouncilarchives.ca/RAD_chap9_revised_Aug2003.pdf.

¹⁰⁹ LAC, *File Format Guidelines for Preservation and Long-term Access Version 1.0*, <http://www.collectionscanada.gc.ca/digital-initiatives/012018-2200-e.html>. (accessed February 2011).

¹¹⁰ InterPARES http://www.interpares.org/ip2/ip2_index.cfm (accessed 6 September 2007).

¹¹¹ IP2 Case Studies, http://www.interpares.org/ip2/ip2_case_studies.cfm (accessed 29 August 2012).

¹¹² IP2, General Study 10 Report and associated documents are available at http://www.interpares.org/ip2/ip2_case_studies.cfm?study=34 (accessed 21 August 2012).

¹¹³ Tracey P. Lauriault and Yvette Hackett, *CS06 Cybercartographic Atlas of Antarctica* (Vancouver, 2005).

¹¹⁴ Kenneth Hawkins, *CS19 Authenticating Engineering Objects for Digital Preservation* (Vancouver, 2005).

¹¹⁵ Jean-François Blanchette, Françoise Banat-Berger and Geneviève Shepherd, *CS18 Computerization of Alsace-Moselle's Land Registry* (Vancouver, 2004).

¹¹⁶ See Table 1: Authenticity Statements in the supplementary file on the IP website at http://www.interpares.org/display_file.cfm?doc=Archivaria64_Todays_Data_supplementary_tables.pdf (accessed 29 August 2012).

¹¹⁷ Trust in data sources was found to be important in both the Cybercartographic Atlas of Antarctica (CS06) and the Archeology records study (CS14), while technological integrity, validity procedures and system checks were implemented in the Mars Global Surveyor Data Study (CS08), the Computerization of Alsace-Moselle's Land Registry (CS18) and the Most Satellite Mission project (CS26). Some of the studies controlled access to their datasets via the appointment of responsible agents (VanMap and the Alsace-Moselle Land Registry), or specialists (Cybercartographic Atlas), while peer

of authenticity in the same way that archivists do, as the emphasis leans more towards measures to ensure data quality. The Land Registry (CS18) may be the exception, as the system is specifically designed to ensure that each registration is authenticated in the system.

Table 5. InterPARES 2 Geospatial Data Case Studies.

Case Study Title	Description	Observational Data	Computational Data
<i>CS06 - Cybercartographic Atlas of Antarctica</i>	Online interactive and dynamic, open standards, interoperable multimedia, multisensory, multimodal atlas that renders distributed data from myriad scientific organizations.	Distributed observational data from myriad sources in real time, film, satellite images, tabular data, sounds, etc.	Data are rendered /refined into maps, charts, tables by the Nunaliit Atlas Framework
<i>CS08 - Mars Global Surveyor Data Records in the Planetary Data System</i>	Surveyor mission data records at the Planetary Data System (PDS) Space Science Data Archive.	Level 0 and refined planetary spacecraft mission data stored in a database with attributes and metadata	Software used to access or visualize the data
<i>CS14 - Coalescent Communities in Arizona</i>	Archaeological Records of the American Southwest rendered in a GIS.	Tabulated raw data collected from archaeological digs	Raw data are rendered/refined into maps
<i>CS18 - Computerization of Alsace-Moselle's Land Registry</i>	Electronic registry including digital transcription of 40 000 existing paper registries and new database entries individually signed by a judge using a PKI infrastructure combining biometric access and digital signatures.	Database of digitized paper land registries and attributes	Data and their attributes are accessed via a data base using PKI technology
<i>CS19 - Authenticating Engineering Objects for Digital Preservation</i>	Examines through an engineering experiment the authentication of digital model (CAD) records using a content/message/semantic-based methodology rather than media, bit-count, or static provenancial attribute-based authentication.	CAD solid model files used in the design and manufacturing of mechanical piece-part assemblies	The files are rendered and stored in a proprietary system
<i>CS24 - City of Vancouver Geographic Information System (VanMap)</i>	An enterprise web-based map system maintained by the City of Vancouver's Information Technology Department.	Land use, social statistics, city infrastructure data both internally collected and acquired from external sources etc.	Enterprise GIS renders data into maps, charts, and tables etc.
<i>CS26 – MOST Satellite Mission: Preservation of Space Telescope Data</i>	Repository of the Microvariability & Oscillations of Stars satellite mission data of Canada's first space telescope.	Raw satellite data and their refined counterparts	

review of data was a method used in the Mars Global Surveyor Data study (CS08). In the Engineering Objects Study (CS19), the creators did not believe their records to be authentic as they have no assurance system.

Portals, it was discovered during the IP2 science case studies, provide a framework within which geospatial data archivists can work and from which they can expand with policies, standards and metadata. Especially since, data producers and users have already appraised the data these contain or refer to as being valuable enough to be paid for, collected, described, licensed, endorsed, organized and disseminated. Also, geospatial data are increasingly being discovered and accessed via data portals. Portals make it possible for users to “gather data germane to their own needs more readily, extract data from online and other electronic repositories, develop the information product they need, use the products for decision making, and contribute their locally gathered geoinformation and derived products to libraries or other repositories.”¹¹⁸

As is commonly known in the field of geomatics, portals can provide all or some of the following services: search and retrieval of data; item descriptions; display services; data processing; the platform to share models and simulations; and the collection and maintenance of data. Much but not all of the data derived from portals are raw in nature and require the user to interpret, analyse and/or manipulate them while the reasons for portal creation are one-stop-shopping, distributed responsibility over data sets, discoverability, and reduction in cost as data are stored or described once and used many times.¹¹⁹

Furthermore, portals are the technical embodiment of data-sharing policies. Individuals within organizations, research projects, or scientific collaborations register their data holdings in portals via an online form organized according to a metadata standard, and then choose to make their data available for free or for sale, viewing or downloading.¹²⁰ Metadata standards “establish the terms and definitions to provide a consistent means to describe the quality and characteristics of geospatial data,”¹²¹ and the ISO 19115 metadata¹²² standard has become an international standard in the field of geomatics as have those mentioned earlier in the case of the IPY. Most of the portals examined in IP2’s General Study 10 include either very detailed metadata or rudimentary header information that contains lineage information.¹²³

The architecture of data portals varies¹²⁴ and portals can be a single enterprise sponsored portal (like a national library), a network of enterprises (like a federation of libraries) or a loose network connected by protocols (like the Web). Distributed data portals have datasets described according to a given standard, and when a request is sent to them by a given site a search is executed by a search agent¹²⁵ to access or render the data into a map or some other form. Many maps and atlases for example, adhere to interoperable OGC standards and specifications, which allow for distributed mapping. In the case of *Cybercartographic Atlas of Antarctica*, for example, users access a module and a call is made to the British Antarctic Survey data portal in the United Kingdom and these are then rendered into a map in real time by the Atlas Framework in Ottawa and delivered directly to the user’s computer. Other examples of

¹¹⁸ US National Research Council, (1999), Spatial Information Resources, *Distributed Geolibraries* (Washington DC), (<http://www.nap.edu/openbook.php?isbn=0309065402>), p. 36. (accessed 29 August 2011).

¹¹⁹ See Tracey P. Lauriault, (2003). A Geospatial Data Infrastructure is an Infrastructure for Sustainable Development in East Timor, (Master’s Thesis, Carleton University).

¹²⁰ Lauriault (2003).

¹²¹ Nancy Tosta and Michael Domaratz. “The U.S. National Spatial Data Infrastructure,” in *Geographic Information Research: Bridging the Atlantic*, ed. Massimo C. Craglia and Helen Couclelis (London: CRC Press, 1997), 22.

¹²² International Standards Organization (2003). *Geographic information – Metadata* (accessed March 2011), (<http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35&ICS2=240&ICS3=70>).

¹²³ See Table 3 Metadata on the InterPARES 2 website at http://www.interpares.org/display_file.cfm?doc=Archivaria64_Todays_Data_supplementary_tables.pdf.

¹²⁴ National Research Council (1999), Spatial Information Resources, *Distributed Geolibraries*.

¹²⁵ National Research Council (1999).

this type of portal are those, which use Web mapping services, such as the British Antarctic Survey. The GeoConnections Discovery Portal (IP2SF15) is an example of this type of portal. Data portals can be housed in a single physical location (e.g., Statistics Canada), and they may be virtual, housed in a set of physical locations and linked electronically to create a single, coherent collection (e.g., GCMD, International Comprehensive Ocean Atmospheric Dataset). The distinction between centralized, distributed or unified portals may have funding, policy and preservation implications.

There are three functional data collections/portal categories¹²⁶ such as 1) research data collections; 2) resource or community data collections; and 3) reference data collections, and each have sets of characteristics that affect the type of data they contain, data contributors, funding and institutional arrangements and the kind of preservation strategies that may apply.

The GS10 portals provided a rich array of information on the topic of authenticity. Ensuring that data are of good quality and can be trusted is critical to these data portals, or else users would not rely on them. For many of the portals, the process of data control begins at the time the data are ingested into the system, made accessible via Web server sharing protocols, or described in a metadata description form. Some portals only ingest data that are derived from peer review journals; others restrict who can access the portal and contribute data.

Once data are in a particular portal, there are a wide variety of security measures in place to ensure they are not tampered with. Before data are made available, most have validation processes in place to attest to their authenticity and quality. However, there did not seem to be any mechanisms for users to assess if the datasets they downloaded or received are authentic beyond metadata and file headers. Also, the concept of the presumption of authenticity defined as “an inference as to the fact of a record’s authenticity that is drawn from known facts about the manner in which that record has been created and maintained”¹²⁷ is a useful one as the context, practices, associated documentation, validation processes and authentication, and access measures would suggest that the Case Studies discussed and portals discussed here are presumed authentic from an archival perspective.

Table 6. General Study 10 Geospatial Data Portals

Portal Title	Description
Canadian Geospatial Data Infrastructure (CGDI) Access Portal	Geospatial data such as DEM, orthophotos, air photos, satellite and radar imagery, data Services, discovery metadata, documents, Internet Maps, and Atlases.
Statistics Canada	Under the Statistics Act Statistics Canada is required to collect, compile, analyse, abstract and publish statistical information relating to the commercial, industrial, financial, social, economic and general activities and conditions of the people of Canada.
Long-term Ecological Research (LTER)	Data of Holocene barrier island geology; salt marsh ecology, geology, and hydrology; ecology/evolution of insular vertebrates; primary/secondary succession; life-form modeling of succession; online maps, photos, webcams; physical; biological, images and geographic data, software, LTER Network Data; and models.

¹²⁶ For detailed descriptions of these see the American Institute of Physics (AIP), 2001, *AIP Study of Multi-Institutional Collaborations: Final Report. Highlights and Project Documentations* Melville, accessed March 2011, <http://www.aip.org/history/publications.html> and National Science Foundation, 2005, *Report of the National Science Board*, p. 20. The GS10 Report describes this more thoroughly and includes portal examples.

¹²⁷ InterPARES 2, *Terminology and Glossary*, http://interpares.org/ip2/ip2_terminology_db.cfm.

Southern California Earthquake Center (SCEC)	Data sets include seismic waveforms, mostly passive source seismic data collected by broadband, strong-motion and analogue instruments. LARSE I and II seismic survey, data sets from the portable deployments following the Landers and Northridge earthquakes, and data from the Anza network and SCEC borehole stations. Non-seismic data includes “survey-mode” precise GPS measurements made in southern California by various universities. The GPS data archive consists of raw GPS data, RINEX files, indices to the RINEX files, log sheets and site descriptions.
International Comprehensive Ocean Atmosphere Data Set (ICOADS)	Surface marine reports from ships, buoys, and other platform types. Each report contains individual observations of meteorological and oceanographic variables, such as sea surface and air temperatures, wind, pressure, humidity, and cloudiness; monthly summary statistics.
National Geophysical Data Center (NGDC - NOAA)	Geophysical data describing the solid earth, marine, and solar-terrestrial environment, as well as earth observations from space.
Antarctic Digital Database (ADD)	1:200,000/1:250,000 maps and collaborative topographic database compiled from a variety of Antarctic map and satellite image sources.
National Snow and Ice Data Center (NSIDC), NASA	Snow, ice and glaciological Data From satellite images remote sensing instruments, ground measurements, and data models.
U.S. Antarctic Resource Center (USARC)	Maps, orthophotos, satellite imagery, point data, Digital Elevation Models, Digital Raster Graphics, and aerial photography.
British Antarctic Survey (BAS) -Antarctic Environmental Data Centre	Antarctic research results, climate modeling and predictions. Also, oceanic, environmental, atmospheric, geoscience, water, and earth observation data from a variety of sensors and in many forms.
Global Change Master Directory (GCMD) –Global Change Data Center	Data inform climate change research, and datasets include models, instruments and services. It is a workspace as well as a place to store these working operational models. Disciplines include atmospheric science, oceanography, ecology, geology, hydrology, and human dimensions of climate change.
Community Data Portal at NCAR	CDP catalogs observational and computer simulation datasets, sources of data are related to sciences in the areas of: oceanic, atmospheric, space weather, and turbulence.
Earth Systems Grid (ESG) portal	Climate change models include High-resolution, long-duration simulations, data simulations, derived from UCAR/NCAR projects (particularly the Community Climate System Model (CCSM) and Parallel Climate Model (PCM) and the Intergovernmental Panel on Climate Change (IPCC). Also physical” datasets that are generated directly by climate simulations.
USGS Data Portals - GEO-DATA Explorer (GEODE)	Geologic discipline’s programs including coastal and marine geology, earth surface dynamics, earthquake hazards, integrated natural resource sciences, mineral resources, National Cooperative Geologic Mapping, volcano hazards. Also scientific and energy related data.

10. Conclusion

The CLI and the 1986 Domesday rescue and salvage examples demonstrated that preservation as an afterthought is uncertain, will most likely yield only partial results, is expensive and time consuming and is often not 100 percent successful. Also, that maps and atlases designed in closed proprietary systems impede preservation and furthermore, that saving the data is not enough, as entire systems either need to be preserved or recreated to regain intended functionality and aesthetics. For every successful recovery there are many datasets, which have been permanently lost, and many more which are at risk. The two examples given were only saved because of their high profile and demonstrable scientific and cultural importance. The three cybercartographic atlases discussed became community specific archives created outside of the traditional archival system but exemplify proactive archiving in their design. In addition, these are making records by recording oral knowledge and they are contesting the validity of official

treaty archival records, the transmediation of ‘official’ archival records into a digitized form and are creating conditions for secondary provenance. They are re-mapping colonial records with indigenous geonarratives and by doing so are re-interpreting old maps and legal processes according to an indigenous worldview. By examining existing Government of Canada geospatial data archiving initiatives and data management models, it was discovered that complex sensor derived geospatial databases are being proactively preserved although none are being ‘officially’ archived in perpetuity by a recognized archival institution and most of the initiatives examined are not governed by a institutional archival or records management policy. Regardless, these are concrete cases upon which to build. With regards to consultations, studies, reports and initiatives, few address geospatial data explicitly although research data findings and recommendations are mostly transferable to a geospatial context and in some cases recommendations from the public on the topic of archiving were ignored. For the last decade substantial studies have one after the other identified the issues and what needs to be done but none of the key recommendations have been implemented. Also, Canadian experts and citizens make valuable contribution when consulted and government continues to commission studies, which in the main seem to gather dust on shelves or in government servers. The reports and studies are accessible but digital data continue to be at increasing risk and many datasets are lost with limited or no, hope of recovery. People, however, remain committed to see data being preserved and show their commitment by participating in Research Strategy Canada, CODATA and the OGC. In terms of legislation, regulation, directives and policies, the creation of government geospatial data are well covered, but these are rarely implemented and there seems to be no implementation mechanism. There are few existing geospatial data preservation initiatives, and the sorry state of LAC and the lack of trusted digital repositories and records management of data overall is of great concern. It was also seen that laws, rules, directions and policy are irrelevant in Canada unless there are dedicated resources set aside to implement them in the government bureaucracy in real terms. In addition, LAC guidelines for the preservation of geospatial data and maps cannot be effectively used by geospatial data managers, and with recent budget cuts, it is uncertain what LAC’s capacity will be to assist data producers with preservation. It would seem, that as an interim strategy, it is best that data producers and managers incorporate life cycle records management into their processes and build in proactive archiving practices. Regarding research in Canada, the IP2 project was groundbreaking with its Case Studies and General Study 10 and these have provided much insight into geospatial record making practices. These studies provided empirical evidence that geospatial data dissemination and map-making processes generally have many good practices in place, most notably metadata, open source technologies and specifications which make it easier for their products to be archived. Furthermore, data in portals have already been appraised by institutions, and archivists can build archival practices into these. It is hoped that IP3 will implement the findings of IP2 for geospatial data and maps. Also, as seen, GeoConnections commissioned a second study by HAL to examine the preservation of its geospatial data assets, and it was argued that the CGDI is a good model within which preservation can be embedded.

Overall, we are losing digital data, maps and atlases faster than we are producing them. There are weak guidelines, technical and institutional infrastructure is sorely lacking, and limited financial resources have been allocated to operationalize laws and regulation. We have also seen glimmers of hope in some initiatives, and there is tremendous will as witnessed in consultations with experts. Archivists have portals upon which they can build, and some of the examined geospatial data preservation initiatives, portals and all the cybercartographic atlases are proactively building in preservation and records management practices.

Geospatial data, maps and atlases as shown in this paper, are a fundamental source in our memory of the world. They form part of our collective memory system, they help us understand our geonarratives, they counter colonial mappings, are the result of scientific endeavours, represent multiple worldviews, and they inform decisions. We simply need to overcome the many challenges preventing their preservation and build upon existing preservation examples and implement our laws, regulations, directives and policies.

Preserving Tradition and Performing Arts in Digital Form

Life at the Edge of the Internet

Preserving the Digital Heritage of Indigenous Cultures

Ravi Katikala,¹ Kurt Madsen² and Gilberto Mincaye Nenquimo Enqueri³

¹Director, OpenText, rkatal@opentext.com; ²Software Architect, MetaTech, kmadsen@metatech.us;

³Vice President, N.A.W.E. / Nationalities Waorani of Ecuador, ngilbertosilva@yahoo.com

Abstract

This paper presents our research and field work with the Waorani Indians in eastern Ecuador regarding how they can preserve their digital heritage and culture on the Internet. We focused on empowering the Waorani to use technology to approach the Internet on their terms: to tell their story, not have their story told, to be independent, not dependent. Using analogies to life in the jungle, we explored issues such as digital self-determination, proprietary file formats, control of material entrusted to cloud service providers, international data import/export, content ownership vs. licensing, and intellectual property. Archival systems are only as valuable as their input data. This data is at risk due to competing economic and legal forces that can adversely influence content, digitization, ownership, and permitted usage. To address this problem, we present an encryption framework that encourages medical tourism to indigenous villages by protecting archived medical data, privacy, and constitutional rights.

Authors

Ravi Katikala is Director, professional services at OpenText.com, a content management company headquartered in Toronto, Canada. He has twenty-five years of experience in management and technical consulting. His expertise is in architecture, systems integration, and process automation. Mr. Katikala holds a bachelor of science degree in electronics and communications from www.nitw.ac.in/nitw.

Kurt Madsen is a software architect and process automation specialist within the healthcare, financial services, telecommunications, and aviation industries. He serves on the board of directors for www.wmnf.org community radio. Mr. Madsen holds a master of science degree in computer science from www.poly.edu and a bachelor of arts degree in economics/international trade from www.rutgers.edu. He has taught twenty courses in computer science and management at www.usf.edu and www.phoenix.edu. His conference publications and presentations are available at www.acm.org and www.ieee.org.

Gilberto Mincaye Nenquimo Enqueri is Vice President of www.nacionalidadwaorani.org, Nacionalidad Waorani del Ecuador, the Waorani community association serving 2,200 indigenous people in eastern Ecuador. He works on their behalf to protect their interests with the government of Ecuador, particularly in the areas of education, healthcare, territory, and community development. He is passionate about preserving the digital heritage of the Waorani.

1. Introduction

Preserving the memory of the world in the digital age requires more than good solutions to the technical challenges of the digital documentary lifecycle (data capture, indexing, archival, and retrieval). Any system of digitization and preservation is only as valuable as its input data (i.e., “garbage-in; garbage-out”). For this reason, one must also consider the competing economic and legal forces that drive digital content creation, acquisition, processing, ownership, distribution, and permitted usage. For example, the constitutions of the authors’ home countries—Ecuador and the United States of America—guarantee citizens some degree of

digital self determination.¹ Yet, in Florida, USA, it is common business practice for doctors to withhold medical data about patients from patients because they do not want to provide patients with evidence that could be used against a doctor in court (in the event that the patient sues the doctor for medical malpractice). This business practice, combined with proprietary data storage formats and disparate repositories of medical data, makes it difficult for patients to independently access a unified view of their own medical history. The evolution of medical information processing systems in the USA have been driven, in large measure, by the economic interests of and laws written for doctors, insurance companies, and lawyers. This challenge is an example of information hiding, one of the patterns of human interaction on the Internet, which is explored in this paper. Such patterns potentially present challenges to the digital preservation of the heritage of indigenous cultures. Private companies and even governments are not immune to the special interests and market forces that create these challenges. The work and leadership of UNESCO is needed to build consensus around solutions that protect the input to digitization and preservation systems.

This paper is based on our work with the Waorani Indians of Eastern Ecuador. Their quest for digital self-determination, including their desire to overcome the challenges to preserving their digital heritage, provides a proto-typical model for indigenous peoples elsewhere. There are still pockets of indigenous people throughout the world who are isolated, geographically, culturally, and economically from mainstream society. In general, indigenous people can take one of several paths when initially contacting the outside world, specifically via the Internet:

- ***Indigenous communities remain in isolation*** – They can chose to remain in isolation, avoiding contact with the outside world thereby protecting their cultures from external influences. In eastern Ecuador, the government has set aside land for the Tagaeri in a region of the Amazon in which all outside contact or travel is forbidden. The indigenous people living there, a clan of the Waorani, are left in total isolation and can continue their traditional ways of life undisturbed.² Similar groups in Peru and Brazil have had little if any contact with the outside world.³
- ***Indigenous communities take their chances by themselves*** – They explore newly discovered access to the Internet from towns on the edge of the Amazon and must rely on commercial vendors for preservation and digitization. This approach to preserving their cultural heritage consists of random successes and failures. There is potential for the Internet and all commercial interests that come with it to dilute indigenous cultures.
- ***Indigenous communities receive guidance*** – This is a hybrid approach in which a community receives guidance with digitization and preservation, generally from non-profit non-governmental organizations (NGOs) whose mission and values are aligned with the best interest of the indigenous community.⁴ This is our recommended approach and the basis of this paper. This is the best chance for indigenous people to maintain their digital self-determination as they preserve their culture, history, and legacy on the world wide web. This path requires the leadership of UNESCO and similar organizations.

¹ See the 4th amendment of the constitution of the USA. http://www.law.cornell.edu/constitution/fourth_amendment. [Note: All on-line references valid as of September 24, 2012.]

² “Territories of Nationality Waorani of Ecuador.” <http://nacionalidadwaorani.org/territorio.html>.

³ J. Scott, “Indigenous Peoples Living in Voluntary Isolation,” United Nations. www.un.org/events/tenstories/06/story.asp?storyID=200.

⁴ <http://www.saveamericasforests.org/Indigenous/indigenous-gallery1.htm>.

Humanity has a choice: for those cases where isolated societies interact with the industrialized world, we can either help them to successfully preserve their cultural heritage on the Internet or let them attempt to make it on their own, risking exploitation, cultural dilution, and loss of digital memory of the world. We see at least three aspects of providing guidance to indigenous communities:

1. ***Legal frameworks that protect the ability for indigenous people to preserve their digital heritage*** – Organizations like UNESCO publish recommendations that governments enact and enforce laws to better protect the digital content acquisition, processing, ownership, distribution, and permitted usage of digital content. This will, in turn, protect the input to digitization and preservation systems. In computer science nomenclature, this problem is known as “garbage in – garbage out.” Any system of digitization and preservation is only as good as its input data.
2. ***Education for indigenous authors regarding the challenges to preserving digital heritage*** – Educate indigenous communities so that they are aware of these challenges. This paper presents our experiences in Ecuador to build common understanding of these challenges as well as potential solutions. This education process is particularly difficult given that indigenous people living in isolation have no prior experience nor frame of reference with which to approach concepts of digitization and preservation in the digital age.
3. ***Technical solutions that mitigate these challenges despite environments with competing commercial interests*** – Technical solutions such as identity escrow achieve a balance between protecting the integrity of information entered into archival systems v. the commercial influences that bias it. It is important protect the digital rights of content originators: both the Waorani in Ecuador and the Musqueam native Americans in British Columbia have independently expressed the same concern, namely their intent to retain some degree of ownership and editorial control over indigenous content when published by third parties. Further, the commercial interests of content originators and the owners of archival systems are not always aligned (e.g., the archival of medical data by health insurance companies in the USA). This paper presents an example of a technical solution in this area.

This paper is organized as follows. First, we describe the experience of living on the edge of the Internet, particularly for the Waorani people in eastern Ecuador. The Waorani provide a model for indigenous people elsewhere in the world who are also living at the edge of the Internet and seek to preserve their digital heritage. Next we explore the import/export of information across the edge of the Internet in the context of constitutional rights and digital self-determination of indigenous communities. We then identify challenges, in the form of patterns, to preserving the digital heritage of indigenous cultures. We show how analogies between patterns in the digital world vs. life in the jungle can provide common understanding when educating indigenous communities on how to best approach the Internet, specifically to document their cultures in digital form. Finally, we present an example of a technical solution that mitigates some of the challenges that they will encounter. We use an example the medical challenges in the USA, which relate directly to challenges the Waorani communities will experience as they archive data captured in their medical clinics in the Amazon. The paper presents an example of how current business practices inadvertently infringe on citizen’s rights to digital self determination, at times even violating constitutional rights. We conclude with recommendations for future work that will help indigenous communities to digitize and preserve their cultural heritage on their own terms.

2. Life at the Edge of the Internet

The expansion of the Internet and its dominant role in our lives has become ubiquitous in almost all corners of the world. Figure 1 shows a map of the Internet that was generated by analysing network traffic volumes.⁵ The lines represent data flowing between the switching nodes in cities. The brighter the line, the higher the volume of data. Stepping back, one can see that this image appears as a network of neurons in a brain. This is a good metaphor for our planet as it becomes wired into a collective consciousness assimilating all peoples and cultures—including isolated tribes in the Amazon.

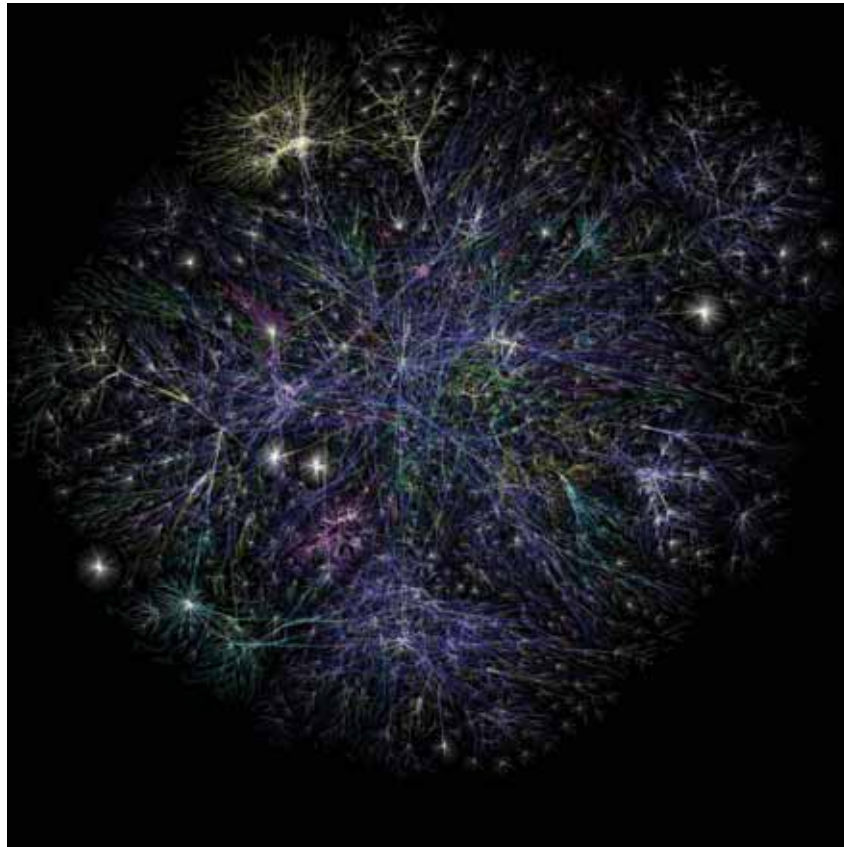


Figure 1. Map of Internet Traffic appears as neurons in the brain.

As humanity's collective consciousness reaches out to the last remaining isolated tribes, we must consider the degree to which our interactions across the edge of the Internet bias the cultures we seek to document and archive. For example within the past century, Christian missionaries have entered eastern Ecuador to spread the Word of God. This has had profound consequences on the local cultures, mostly positive. Internal warfare ceased as each longhouse group of Waorani was contacted by missionaries, beginning in

⁵ <http://www.opte.org/maps>

1958.⁶ On one hand, the tribes living there have stopped killing each other and now live in relative peace. On the other, some traditional ways such as the medicinal treatments of shamans or witch doctors have been suppressed. While much research has been done regarding the cause of these trends, there is no doubt that western influences have the potential to profoundly alter indigenous cultures.

One must visit the Amazon to truly appreciate its vastness and the isolation of its jungle settlements. Many of the roads in eastern Ecuador lead only to the oil fields and logging camps along the edge of the jungle. The only way into or out of isolated indigenous villages is via boat, small aircraft, or on foot. To reach Tzapino, the Waorani village that was studied for this paper, one must fly north east from Puyo, a small city on the edge of the jungle. During the flight, the volcanoes of the mighty Andes rise to the west while the endless expanse of jungle stretches 3,000 kilometers to the east. Upon reaching Tzapino, a skilled pilot must land a heavily loaded airplane onto a 500 meter grass airstrip next to one of the countless small rivers that feed the Amazon. The airfield is deep in the jungle surrounded by trees and traditional Waorani houses made of bamboo and palm leaves. There is no electricity in Tzapino, no Internet, and no running water. It is a day's walk to the nearest dirt road. Communication with the outside world is batched, not real-time. Figure 2 shows a Cessna 206, which is typical of the aircraft used by missionaries and NGOs to fly supplies into the jungle, export handicrafts to markets, and occasionally to evacuate people for emergency medical care (e.g., snake bite).

The Waorani territories in Ecuador cover 790,000 hectares in the provinces of Pastaza, Napo, and Orellana (which includes the Yasuni biosphere). The 2,200 indigenous Waorani who live in these territories are represented by N.A.W.E., the Nacionalidad Waorani del Ecuador.⁷ N.A.W.E. represents



Figure 2. Mission aircraft used to reach indigenous communities.

⁶ Stephen Beckerman, Pamela I. Erickson, James Yost, Jhanira Regalado, Lilia Jaramillo, Corey Sparks, Moises Iromenga, and Kathryn Long, "Life Histories, Blood Revenge, and Reproductive Success Among the Waorani of Ecuador," *Proceedings of the National Academy of Sciences of the USA*, 106, no. 20 (May 2009): 8134-8139, p. 8135.

⁷ www.nacionalidadwaorani.org

their collective bargaining via meetings the government and private companies such as oil drilling and timber companies. The challenges N.A.W.E. and the Waorani communities face are foreign to many in the industrialized nations, and yet the preservation of the digital heritage of these cultures begins with basic necessities such as access to clean drinking water and emergency medical support.

The Waorani now live on both sides of the Edge of the Internet: they retain their jungle communities while also maintaining a presence in frontier towns and cities such as Puyo, Ecuador, population 25,000. Since joining the modern world, the Waorani no longer depend entirely on the natural work of the forest to survive, and they are no longer nomadic. Instead, the young Waorani who live in neighboring towns are concerned with finding jobs, participating in government projects, and of course western culture, particularly as it appears to them on the Internet. This is a concern to the older generations of Waorani who seek to preserve traditional ways and who still view the land as their mother giver of food. To strike a balance, they are seeking technologies that can improve their management model for governance to help their people avoid poverty. They also want to protect their culture, stories, photos, and intellectual property, which is often published by third parties without their permission.

N.A.W.E. has established goals for the Waorani people in four areas: education, healthcare, territory, and community development (undertaken in partnership with the government and NGOs). Our work focuses on the information archival aspects of these goals, particularly education, healthcare, and the multi-jurisdictional legal framework for information exchange across the edge of the Internet.

2.1 Education

Education is not only essential to bring general awareness and knowledge to Waorani children, it is essential for the digital preservation of their culture. The first schools to bring knowledge of the outside world to the Amazon were established by visiting missionaries. Today, the Waorani are exploring Moodle,⁸ a Learning Management System, to enhance the learning process, including distance learning. Due to the isolation of many Waorani communities—specifically the lack of electricity, telecommunications, and Internet infrastructure—indigenous students will use solar-powered laptops to take off-line classes taught by teachers in Quito, the capital of Ecuador hundreds of kilometers away. Teachers and students will communicate in weekly batch cycles with homework and assignments exchanged via USB flash drives on aircraft serving missionaries, supply lines, and tourism. During this process, the digital heritage of the Waorani can be captured, preserved, and shared with the rest of the world.

Waorani children are curious and intelligent. Even with limited educational facilities, they show tremendous promise. These children love the forest and all of the wonders it brings. The forest is part of their heritage. As they cross the edge of the digital divide, they will be able to make significant contributions to the memory of the world by sharing the secrets of Amazon.

2.2 Healthcare

Many of the Waorani do not have access to modern medical facilities in their jungle communities. In the event of an emergency (e.g., snake bite) or a serious illness, a patient must be transported to the nearest city for treatment. Often times, the availability of transportation makes all the difference between life and

⁸ www.moodle.org

death. These communities need medical clinics to serve local residents as well as medical tourists from other countries (most notably the USA) seeking alternative medicines found only in the Amazon.

These Indian tribes for the most part have lived in isolation from other civilizations, and they have never been exposed to diseases, like Influenza A, which are very common in other parts of the world. These communities have no significant cases of heart disease, cancer, or stroke. Their unique immune system and dietary conditions provide scientific clues into the origins of certain diseases. Dr. James W. Larrick, a physician who made several trips to Waorani community during 1976 conducted several studies on relatively pristine immunologic state of these tribes.⁹

From these perspectives, the Waorani are prototypical of indigenous communities elsewhere in the world. They provide a good example for studying how to preserve the digital heritage of indigenous cultures.

3. The Amazon Information Pipeline

The first aspect of guiding indigenous communities towards digital self-determination is to provide them with an understanding of the legal frameworks that govern their ability to preserve their digital heritage.

Fulfilling Waorani goals—in education, health, territory, and community development projects—involves the free flow of information between the Waorani and other parties such as publishers, research scientists who visit the Amazon, and international organizations concerned with helping the Waorani (e.g., www.saveamericasforests.org). The Waorani are very sensitive to the extraction—and in some cases exploitation—of natural resources from the Amazon (e.g., oil and timber). Regarding the import/export of their information across the edge of the Internet, they would like to retain some degree of ownership and editorial control over indigenous content when published by third parties. We call this effort iPEACE—Information Pipeline for Education and Amazonian Culture Exchange.

3.1 Importing Information into Indigenous Communities

Importing information from industrialized world into Waorani communities appears to pose little legal risk—the Ecuadorian constitution guarantees the right to free access of information. However, once an indigenous community starts communicating and interacting independently with outside cultures, they enter legal structures and economic interests that might intentionally or unintentionally exploit their culture. The following are some excerpts from the Ecuadorian Constitution regarding the [digital] rights of Ecuadorian Citizens:¹⁰

- Free, intercultural, inclusive, diverse and participatory communication in all spheres of social interaction, by any means or form, in their own language and with their own symbols.
- Universal access to information and communication technologies.
- To build and uphold their own cultural identity, to disseminate their own cultural expressions, and to have access to diverse cultural expressions.

⁹ Richard D. Lyons, “A Doctor in the Amazon Probes for Genetic Links to Disease,” *The New York Times*, November 8, 1983. <http://www.nytimes.com/1983/11/08/science/a-doctor-in-the-amazon-probes-for-genetic-links-to-disease.html>.

¹⁰ http://www.mmrree.gob.ec/pol_exterior/constit_eng.pdf.

- The right to protection of personal information, including access to and decision about information and data of this nature, as well as its corresponding protection. The gathering, filing, processing, distribution or dissemination of these data or information shall require authorization from the holder or a court order.
- The right to inviolability and secrecy of hard-copy and on-line correspondence, which cannot be retained, opened or examined, except in those cases provided by law, after court order and under the obligation to uphold the confidentiality of matters other than those motivating their examination. This right protects any type or form of communication.

The Ecuadorian constitution confers additional rights to Indigenous Communities and Peoples:

- All forms of appropriation of their knowledge, innovations, and practices are forbidden.
- To uphold and develop contacts, ties and cooperation with other peoples, especially those that are divided by international borders.
- Ecuador is the first country to recognize Rights of Nature in its Constitution.¹¹

Duties of the state:

- Facilitate the creation and strengthening of public, private and community media, as well as universal access to information and communication technologies, especially for persons and community groups that do not have this access or have only limited access to them.
- Not permit the oligopolistic or monopolistic ownership, whether direct or indirect, of the media and use of frequencies.

3.2 Exporting Information From Indigenous Communities

Special care should be taken when introducing an indigenous community to the Internet for the purpose of bringing that community's culture to the outside world. Protecting indigenous culture through western intellectual property protection systems is difficult, particularly when copyrighting folklore. The international systems of copyright protection require elements not traditionally found in folklore, such as individual ownership, fixing the copyrighted material to a fixed tangible medium, and minimum standards of individual creativity. Should a native story be incorporated into a book or film, the Waorani want to have copyright protection to ensure that the story is presented accurately. Although the Ecuadorian Constitution guarantees such rights to local communities, a closer look into international enforcement and Ecuador's political position is essential to determining protection. When introducing the culture to the Internet or bringing the community's culture to the outside world, care must be taken so that any profit generated from their contributions are properly distributed to the rightful owners. The current global intellectual property protection systems may have to be extended or modified to cover the special nature of information shared by these communities such as their folklore or medicinal methods.

Furthermore, when western pharmaceutical companies extract information from the Amazon regarding indigenous plants and remedies, their focus is on economic interests such as securing patents; less consideration is given to preserving the culture and stories that accompany the use and application of those plants and traditional remedies, even when this accompanying cultural content is captured.

¹¹ <http://therightsofnature.org/ecuador-rights/>.

4. Education – Challenges and Awareness

Education is the second aspect of guiding indigenous communities towards digital self-determination as they preserve their digital heritage. Education empowers them to deal with challenges on their own terms, particularly in an environment of competing economic forces.

We studied the challenges that they face (when preserving their digital heritage) in the form of patterns of human and business interactions, particularly on the Internet. Many of these concepts are especially challenging because they are both difficult to mitigate and difficult to explain to indigenous authors seeking to document their culture and heritage in electronic form. We found that these patterns and challenges are best explained by analogy to concepts that indigenous people find in the jungle.

- ***Symbiosis (Win–Win)*** – Leaf-cutter ants and the fungus that eats the leaves cut by the ants share a symbiotic relationship. The ants bring leaves back to their nests to feed the fungus that grows in their ant mounds. In turn, the fungus consumes the leaves and produce a by-product that feeds the ants. Neither species can exist without the other. This symbiosis is analogous to the relationship between participants in a digital marketplace: Google’s Android apps and Apple’s iStore are market makers that provide an infrastructure on which market participants build and sell applications. Thus, developing mobile apps that provide access to indigenous content is win-win for everyone.
- ***Framework or Infrastructure (Win–Neutral)*** – Bromeliads are epiphytes that grow on trees; the trees provide the structure and environment in which bromeliads grow, but the bromeliads offer little to the trees. Neither species harms the other. Another example well-known to indigenous communities is a jungle airstrip. By analogy, there are many examples of frameworks and infrastructure in software. One of them is Microsoft’s .NET.
- ***Viruses in Software (Win–Lose)*** – Strangler figs are vines that grow around trees, constrict them, and eventually kill the very hosts that support them. Likewise, computer software viruses infect a host application, cause damage, and eventually die with the rest of the system
- ***Censorship*** – Consider an oil drilling company in the Amazon that requires its employees and indigenous people to sign nondisclosure agreements regarding pollution. How do indigenous communities balance their need for clean drinking water vs. continued payments for drilling rights in their territories? How can they turn to the industrialized world to provide best practices for preserving digital heritage when oil drilling companies in the USA use nondisclosure agreements and sealed court records to block independent scientific research into the adverse health effects of hydro-fracturing drilling?¹² Censorship—however justified—impedes the free flow of information into archival systems, which in turn limits the value of such systems.
- ***Adaptation (or extensibility)*** – Termites adapt to their environment by building nests to suite their environment. In Africa, termites build towers on the ground in a dry environment. In the Amazon, termites build nests in trees so they do not wash away during torrential rainstorms. This adaptability of termites is analogous to the software design concept of adaptability and extensibility, namely that a system (its structure, functionality, and/or intended use) is permitted

¹² Michelle Bamberger and Robert E. Oswald, “Impacts of Gas Drilling on Human and Animal Health.” *NEW SOLUTIONS: A Journal of Environmental and Occupational Health Policy* 22, no. 1 (2012): 51-77.
<http://baywood.metapress.com/openurl.asp?genre=article&id=doi:10.2190/NS.22.1.e>.

to change over time, subject to constraints such as technical limitations or contractual obligations to customers.

- **Information hiding** – A publisher blocks, limits, or alters information produced by a content originator (both roles may be the same entity). Example one (positive): Digital Rights Management (DRM) systems use encryption to protect the intellectual property rights of the content owners by hiding information from non-authorized parties. Example two (negative): In the USA, doctors withhold information from their patients to protect against being sued. In some cases, patients withhold information from their doctors to protect their privacy (because doctors are required to release medical data that they collect to insurance companies).
- **Information Expiration** – Information in an archival system is destroyed after a certain period of time, generally to save disk space once records retention requirements have been met. This pattern is analogous to the spoiling of food in the jungle. Another example is store receipts with invisible ink that disappears after the merchandise return period has expired (the Waorani do see receipts with invisible ink at stores in the town of Puyo). Another example is a doctor who uploads information with the intent that it will only be available for a limited time.
- **Proprietary vs. open-source data formats and computer software** – It is important that indigenous authors understand both the licensing terms and encoding schemes of the software they select record and archive their cultural heritage. Proprietary software is typically leased as a license to access, but not own. Open-source software is generally distributed freely with support services sold separately. To explain these concepts, we demonstrated proprietary vs. open file formats to the Waorani leadership. We showed them the same file as viewed in both Microsoft Word 2007 running on Windows 7 and LibreOffice running on Ubuntu. Then we showed them how these two programs store files in different file formats. It became clear to them that open formats such as DocBook XML can be viewed independently of the source application while proprietary formats such as Microsoft Word require the original program for interpretation.

Figure 3 maps these patterns and metaphors to concepts in the jungle. Additional patterns follow:

Amazon Metaphors for Education (Indigenous View)										
Internet Patterns (Western View)		<div> <div>Termite</div> <div>Leaf-cutter ants</div> <div>Broniads</div> <div>Strangler figs</div> <div>Inter-tribal negotiation</div> <div>Jungle runway</div> <div>Medical Tourism</div> <div>Inter-tribal negotiation</div> <div>Oil company land deals</div> </div>								
	Symbiosis	E/B								LEGEND (regarding protecting indigenous cultures) L = Legal \$ = Economic T = Technical E = Educational ----- B = Beneficial H = Harmful N = Neutral S = Solutions exist
	Framework or Infrastructure	E/B			\$/B					
	Virus in Software (Win-Lose)		T/H							
	Censorship		L/H							
	Adaptation (or extensibility)	E/S								
	Information hiding (patient's view)				T/B					
	Information Expiration (Doctor's view)				T/B					
	Proprietary vs. open source		T/H							
	Broker							L,S/B		
	Information escrow (separation of concerns)						L,S/S			
	Monopoly of supply							L,S/H		
	Monopoly of demand							L,S/H		
	Value-added reseller							\$/B,S		
	Affinity groups									
	Social network (mesh)									

Figure 3. Guiding the Waorani towards digital self-determination with metaphors.

- **Intellectual property, leasing vs. owning** – if the Waorani post their stories on the Internet, who owns them? How are they paid for their content? Do they sell a subscription or make a one-time sale of their stories? Indigenous people should understand the consequences of this issue prior to signing contracts involving intellectual property ownership.
- **Intellectual property, exclusivity** – in intellectual property law, exclusivity applies when the subject of content grants exclusive rights to a single party (such as a publisher) and foregoes the possibility of granting similar rights in the same content to another, unrelated party. We recommend that indigenous communities grant non-exclusive rights to their stories. This way, they retain ownership over their content and the ability to resell or re-license it to another party in the future.
- **Hyperlink Transitivity** – the legal claim that a website owner becomes responsible for all content reachable by hyperlinks on the site. If an Internet user visits a Waorani website, and that site links to other blogs (e.g., perhaps critical of the oil industry), do the Waorani become liable for third party blogs?
- **Content aggregation** – Mixed content from different sources, in particular different indigenous communities that aggregate their content together to make a compelling website about their region.
- **Broker** – A neutral, third party who mediates between a provider or a consumer. For example, a real estate agent takes a commission to broker sales. An electronic clearing house of data or transactions that charges an e-fee to connect indigenous people (e.g., to bring Indian handicrafts to market).
- **Information Escrow** – A variation of the broker pattern in which an independent, third party holds information on behalf of two counter-parties until a transaction between those two counter-parties has been completed. For example, if a publisher insists that Waorani stories be encoded using a proprietary digital rights format, then the Waorani could insist that the publisher keep the original content in unencrypted form on deposit with an escrow service.
- **Information quarantine** – Temporarily isolating data or programs so that they cannot interact with the rest of a computer system. For example, antivirus software quarantines suspicious files.
- **Monopoly of supply** – one company (monopoly) or a few companies (oligopoly) that dominate a market by controlling the supply of product or server. This is the classic definition of monopolistic power. The Organization of Petroleum Exporting Countries (OPEC) is an example of an oligopoly because it controls a large portion of the world's supply of oil. The indigenous people of the Amazon understand the power of oil drilling companies.
- **Monopoly of demand** – one or a few companies dominate a market by controlling the demand for a product or service, or the payment thereof. The health insurance companies in the USA provide an example by effectively acting as one organization by sharing private patient data and controlling almost all payments to medical providers such as doctors (i.e., the demand-side of the monopoly). Unfortunately in the USA, insurance companies are except from the Sherman Anti-Trust Act, a federal law indented to protect citizens against the abuses by monopolies.¹³ Rescinding this exception would improve healthcare in the USA.

¹³ http://en.wikipedia.org/wiki/Sherman_Anti-Trust_Act; http://www.huffingtonpost.com/dylan-ratigan/why-would-we-let-them-rig_b_302480.html.

- **Value-added reseller** – A travel agent who’s website sells tour packages into the Amazon and provides additional services to complement those of the Waorani.
- **Affinity groups / social networks** – The ability for like-minded people to gather on the Internet for a common purpose. For example, Mycologists traveling to Waorani territory study mushrooms and fungus. An affinity group would allow Mycologists to interact with each other and share data on the Internet—under terms controlled by the Waorani—for their common causes such as researching mushrooms within Waorani territory. Another example is enabling the Waorani to band together to face the oil companies with one e-voice.
- **Lock-out** – Some doctors in the USA do not take cash (even though US currency states that is “legal tender for all debts public and private”). Instead, these doctors only see patients who can pay for services through health insurance. Further, if a patient has insurance, some doctors will not allow that patient to pay with cash.
- **Extortion (a.k.a. Protection)** – A third party demands payments for continued business operations by threatening to shutdown the victim’s website. Criminals flood the website with fake page hits to flood the servers in a denial of service attack. Service is restored when the victim pays “protection” money.
- **Reuse by framework** – A software framework is a reusable infrastructure that can be extended and customized to solve similar problems.
- **Reuse by libraries** – A collection of software building blocks that can be combined to solve similar problems.
- **Protocols** – peer-to-peer, point-to-point, multi-cast, push vs. pull
- **Syndicated content** – the ability of the Waorani to join other indigenous communities throughout the Amazon, either to sell their web content (and generate revenue for Waorani communities) or to purchase access to a wider market.
- **Content authentication** – Methods to verify that the content originator is in fact who he claims to be.
- **Content non-repudiation** – Methods to assert that the content originator cannot deny previously originated content.
- **Copyright law varies by country** – Copyright laws protecting Waorani content vary from one country to the next.

5. Technical Solutions to Preserving Digital Heritage

The legal frameworks that govern the preservation of digital heritage often come into conflict with competing commercial interests that drive digital content processing, ownership, and permitted usage. Technical solutions exist to mitigate these challenges. Providing such solutions is the third aspect of guiding indigenous communities towards digital self-determination as they preserve their digital heritage.

A key theme of this paper is that archival systems are only as valuable as their input data. A secondary theme is that competing economic interests may bias, impede, or even prevent the free flow of

information into archival systems. Thus, organizations like UNESCO can help indigenous societies by bringing awareness to research into technical solutions that protect indigenous rights to digitally preserve their culture and heritage in an environment of competing commercial interests.

For example, the Waorani would like to use the Internet and archival technology to preserve and share their ways of traditional medical care using natural resources found in the Amazon. One potential source of this information is Waorani healthcare clinics, which serve the Waorani as well as foreigners—medical tourists—who travel to Waorani territory to seek out traditional healthcare. The Waorani encourage foreigners to visit their healthcare clinics for two reasons: first this brings money into Waorani communities; second, this is another way in which the Waorani can share their cultural heritage and traditional methods of healthcare with the outside world.

Unfortunately, medical tourists from the USA who travel to Waorani territories may be reluctant to visit indigenous medical clinics because some medicines and medical practices from the Amazon may be considered experimental or otherwise not viewed favorably by health insurance companies back home in the USA. Further, the medical data collected from these traditional treatments could be used against them to deny future insurance applications and/or claims. The unintended consequence is a disincentive to archive valuable information. This adversely affects everyone from patients to independent researchers who study anonymous medical data to determine the efficacy of traditional Waorani medicines.

The problem is that in the USA, patients have little control over the scope of information permanently released by their doctors to third parties such as health insurance companies and the Medical Information Bureau (www.mib.com) which archive medical data. These companies do have a legitimate business need to use archived medical data to combat insurance fraud. However, in some scenarios, they can infringe upon a patient's right to privacy by collecting medical data when the patient has no legal obligation to provide it. Consider that the MIB maintains personal medical records for seven years while some insurance applications only require applicants to provide medical history within the past five years. What happens if a person applies for insurance six years after he is diagnosed with cancer? If that diagnosis occurred in the USA, then it would likely be discoverable by the insurance company and used as justification to deny the application for medical coverage. However, if that diagnosis occurred in an indigenous medical clinic outside of the USA, then the insurance company would likely not have access to that diagnosis and would approve the application for medical coverage. In both cases, the applicant can be truthful on his insurance application and no laws are broken. Yet, the outcomes are very different.

The fourth amendment to the constitution of the USA guarantees a citizen's right privacy in their personal effects, which applies in this case. In practice, however, there is little privacy in the USA because businesses control many aspects of the information systems that facilitate daily life in the digital age. The Ecuadorian constitution provides even more explicit protection of personal [electronic] information than the constitution of the USA. But if ubiquitous business practices have the unintended consequence of undermining the constitution of the USA, what is to stop the same fate in Ecuador?

There is a simple and effective technical solution that addresses these challenges via identity escrow, which separates medical data from the authentication data used to identify patients. This separation of concerns gives patients (and potential medical tourists to Waorani territories) the privacy and peace of mind they require before they consent to the collection, archival, and limited, directed release of their medical data. Likewise, from the perspective of medical service providers, the government, and authorized third parties, identity escrow provides the accountability and transparency infrastructure that could support the involuntary release (from the patient's point of view) of authentication and medical data when

necessary to comply with the law. This combination protects the digital self-determination of citizens who provide content for archival systems while meeting the needs of commercial interests.

The rest of this section describes our proposed technical solution, Identity Escrow, in more detail. The system separates authentication data from medication data via two, encrypted databases each with separate access security. These databases are joined together by Medical Record Locator (MRL). The identity database is populated when patients initially register to use this cloud service (e.g., name, address, date of birth, government-issued identification number). The medical database is populated when doctors upload medical data via the MRL after treating a patient. The doctor only has access to the patient's MRL and all associated medical data; no patient-related information from the identity database is available to the doctor. Figure 4 illustrates how the use cases of this technical solution work together.

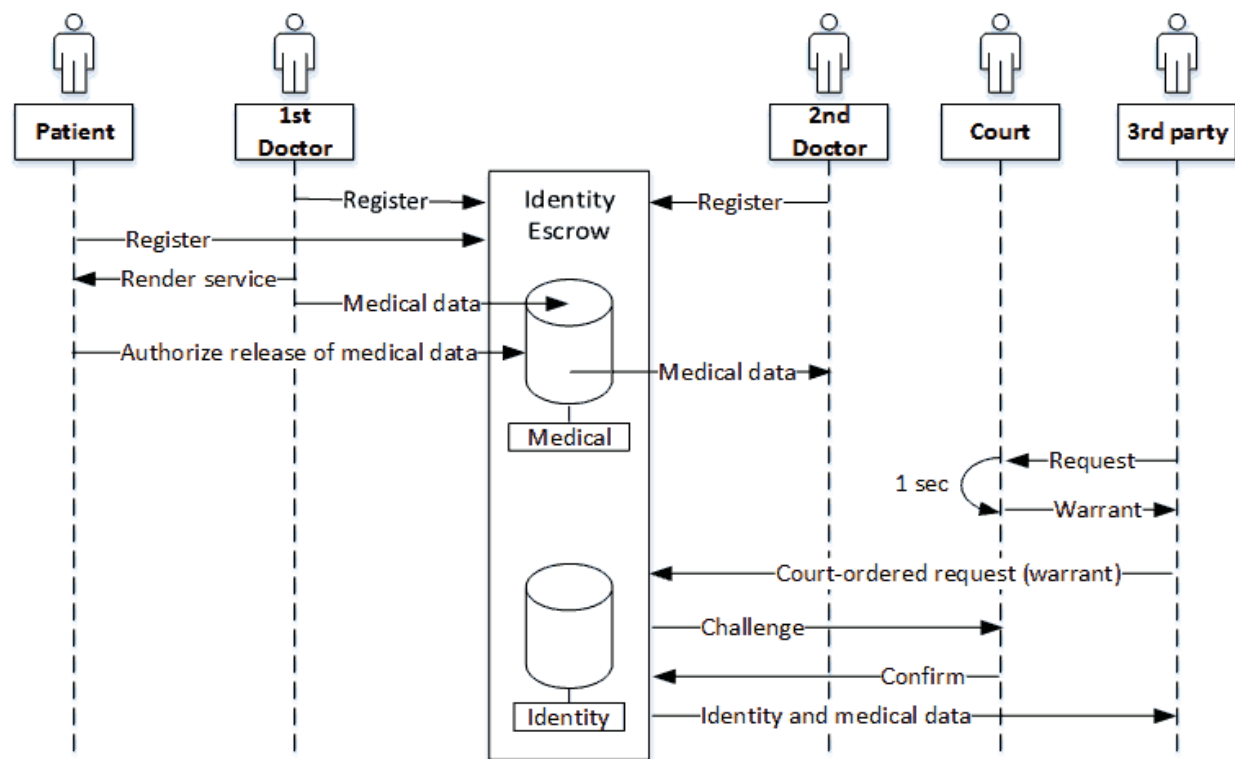


Figure 4. Identity Escrow – Protecting the input data to an archival system from competing interests.

Use case: Medical Service Provider Registration – A provider such as a doctor registers with Identity Escrow by creating an account and providing his or her authenticating information. Depending upon the jurisdiction, it may be necessary for third party such as a government agency to verify that the doctor has a valid license to provided the services in question.

Use case: New Patient Registration – A person seeking medical care with privacy registers with Identity Escrow by creating an account and providing authenticating information. Depending upon the jurisdiction, it may be necessary for a third party such as a notary public to attest that the registration and identity information has been independently verified (e.g., using the patient's passport as documentation).

Use Case: Patient Visits a Provider – When a patient visits a provider such as a medical clinic staffed by doctors, the patient does not provide any identifying information. Instead, the patient presents an Identity

Escrow account number, which the clinic staff use to look up the patient's account. The patient then answers challenge questions, presented by Identity Escrow, to authenticate himself (e.g., to confirm that the account number has not been stolen). Thereafter, the medical examination can commence. Doctors have access to any previous patient medical history provided that the patient has marked such information as visible in his Identity Escrow account. Once the medical services have been rendered, the doctor can upload new medical data into the patient's medical records at Identity Escrow. Payment for services rendered is arranged separately and is outside the scope of Identity Escrow.

Use case: Voluntary release of patient information – A patient may wish to voluntarily release his medical information to a third party, for example to support an application for insurance coverage or to apply for a new job. This use case involves two sub-cases that both start when the patient logs into Identity Escrow to schedule a release of information:

1. The patient identifies a third party who has previously registered with Identity Escrow (e.g., another doctor in the USA who will examine the medical data uploaded at a Waorani clinic in order to render a second opinion). Identity Escrow sends a message with a security key to the third party Identity Escrow with an authentication token that matches information sent independently by the patient to the third party. The key is valid until its time-to-live value expires. Finally, the third party uses the key to pull the required medical data from Identity Escrow.
2. The patient logs into Identity Escrow to schedule a release of information, and Identity Escrow provides the user with a temporary access number. The patient provides this access number to the third party. The third party pulls the relevant medical data from Identity Escrow.

Use case: Involuntary release of patient information – To protect the medical provider and third parties such as insurance companies, Identity Escrow may release protected patient information when necessary to comply with the law. The patient must agree to this condition during the registration process. Generally, a court-ordered warrant or subpoena is required to release authentication and medical data without the patient's consent. This varies by jurisdiction.

Use case: Research – Depending upon the jurisdiction, Identity Escrow may release medical data only (without identifying information) for educational and research purposes.

Other use cases – Other cases include account closure. When a user closes his account, Identity Escrow will maintain archived medical data to comply with requests for involuntary release of patient data.

In summary, Identity Escrow provides an example of how a technical solution can mitigate the challenges to digitization by separating competing concerns, namely the digital content vs. the business transaction that produced it. This protects the content originator's right to preserve his or her digital content in an environment of competing commercial interests. Thus, a content originator—such as a medical tourist to a Waorani alternative health clinic—is encouraged (not discouraged) to authorize the archival of sensitive personal information. The data collected from thousands of patients over time will become part of the digital heritage of traditional Waorani medical treatments. The example presented above mitigates the following patterns of challenges to preserving digital heritage: censorship, monopoly (of demand), information hiding, expiration, and Identity Escrow.

6. Conclusion

We finish with the following conclusions and recommendations for future work:

- There are three paths to preserving the digital and digitized heritage of indigenous cultures: leave them alone, let them take their chances on their own, or provide them with guidance. Guided preservation is the best path, especially when performed by organizations like UNESCO and NGOs whose mission and values are aligned with the best interests of indigenous societies.
- Any system of archival is only as valuable as its input data. This presents a problem when the commercial interests of the owners of archival systems are not aligned with the interests of the content originators. For example, the archival of medical data by health insurance companies in the USA is driven by financial concerns, not the patient's desire to store a life-time portfolio of useful medical information in the context of privacy. This struggle by citizens for digital self-determination is part of on-line culture in the digital age. And it biases the information to be archived.
- Indigenous communities in the Amazon will encounter the same problems as their North American counterparts when they cross the edge of the internet unless organizations like UNESCO and NGOs encourage governments to improve laws that protect individual rights to the digitization and preservation of personal information. Input data that is entered into archival systems depends on these protections.
- Such legal and technical protections will encourage more medical tourists to visit medical clinics in the Amazon to experience alternative medicines and exchange culture with indigenous communities such as the Waorani.
- Technical solutions can mitigate legal, ethical, and economic challenges to digitization and preservation. For example, identity escrow separates the concerns of content originators vs. the owners of commercial archival systems by separating the information produced by a transaction from the financial and personally-identifiable information about the transaction. This facilitates data privacy, right to oblivion, and right to access.
- Indigenous people must develop the skills to approach the Internet on their own terms. This requires education regarding the challenges to preserving their digital heritage. They should not become dependent upon outside influences to control their data and Internet access. Ironically, they need help to do this, namely from trusted organizations like UNESCO that do not have a vested interest in exploiting these societies. This is the surest way that they can empower themselves and preserve their culture in the brave new world. Indigenous information must flow both ways without diluting cultures. They are eager to learn our ways; we must help them preserve theirs.
- Finally, any solution computer-based learning center deployed in the jungle must be self-sustaining and permanent. The indigenous population must be taught to administer their computers by themselves.

Digital Madness, Archival Theory and the Endangered Sound Archives of Radio Botswana

Lekoko Kenosi

Archival Studies Program, University of Botswana

Abstract

In recognition of the contributions of sound archives to the collective wisdom of humanity UNESCO has set aside 27 October as a day for the celebration of Audiovisual Heritage. And despite the shared universal definition of a record as any “document regardless of medium,” the trust placed on paper at the exclusion of other information carriers continues to dominate the mindset of many in Africa. The result of this bias has been little attention paid on the preservation of sound archives. At a professional level, while many theoretical assumptions have been made about the character and treatment of paper records, the literature on the preservation of sound archives continues to come from thinkers without a “mainstream” archival degree. Using the sound archives of Radio Botswana as a case study, this author applies traditional archival theory methods to this genre of records. By so doing the presenter hopes to consolidate, even critique, the application of traditional archival theory to sound archives. In the process of doing that, and with the theme of the conference in mind, an international call to halt the looming eclipse of Radio Botswana sound archives- in an age of digital madness- will be globally launched.

Author

Dr. Lekoko Kenosi has a Bachelor's degree from the University of Botswana, a Master's degree in Archival Studies from the University of British Columbia, Vancouver, Canada and a Ph.D. in Information Science from the University of Pittsburgh, USA. A U.S. Fulbright Scholar in Archival Studies, and a member of the UBC InterPARES project, Dr. Kenosi has published extensively in the area of Records and Archives. He teaches archival theory, records management, electronic records and sound and cinema/television archives at the University of Botswana.

1. Introduction

Even though the first record produced by humans is a sound we take sound so much for granted. In fact the perception that audiovisual records are only good for entertainment is still pervasive. However, that perception is fast disappearing as more archival centers recognize the need to preserve sound archives. In the last few years there has been a proliferation of professional bodies dedicated to the proper preservation of sound archives. Among them Association of Moving Image Archivists (AMIA); Association for Recorded Sound Collections (ARSC); International Federation of Film Archives (FIAPF); Fédération Internationale des Archives de Télévision / International Federation of Television Archives (FIAT/IFTA); Federation of Commercial Audiovisual Libraries International (FOCAL); International Association of Sound and Audiovisual Archives (IASA); SouthEast Asia & Pacific Audiovisual Archives Association (SEAPAVAA) and the Co-ordinating Council of Audiovisual Archives Associations (CCAAA). Even more interesting is the fact that the International Council on Archives (ICA), a global professional association of all archivists, is affiliated to CCAAA, rather than the other way round. Moreover, the terminology or jargon used by these audiovisual associations sometimes differs from that learnt in traditional archival setups and was one of the reasons that motivated this study.

Sound or voice is one of the most powerful means of conveying facts and actions by both literate and illiterate communities. The use of sound to communicate is not just a monopoly of humans only. Domestic and wild animals also communicate their emotions, and maybe their facts and their acts through sound. In its annual commemoration of the World Day for Audiovisual Heritage¹ UNESCO has underscored the importance of sound archives by reminding us that, the first moonwalk original recording that took place on 20 July 1969, cannot be located and is presumed lost. In Africa in general, and in Sub-Saharan Africa in particular, song, dance, oral traditions, folklore, together with the modern written record, are still the chief means of transmitting all forms of knowledge from one generation to the next.

Despite this fact archival theory has not engaged sound records with the same vigor that it has done with the written and the electronic record. Using the sound of Radio Botswana as a case study this research attempted to investigate the application of key components of archival theory to the sound archives of Radio Botswana, drawing in the process strength from Heather MacNeil, who marshaled us to re-examine archival theory”² as a way to challenge old truths. So, what are these components? These key components include:

- Acquisition
- Provenance, Respect des Fonds and Original Order
- Appraisal
- Arrangement and Description
- Access, Preservation and Digitization

In choosing these components I am alive to the fact I am opening myself up to a plethora of criticism. I am back to Vancouver today, to welcome the criticism because archival theory, like all theories, has not developed without controversy. Indeed, a theory that invites passive and monolithic responses is sterile, bankrupt and over time condemns itself to extinction. The above components might sound like a monotonous script to postmodern scholars. In “Archives, Records and Power...” Terry Cook and Joan Schwartz are fatigued by what they call “a script that has been naturalized by the routine repetition of past practice.”³ What Hugh Taylor has referred to as “archival fundamentalism or the reluctance to explore new areas of theoretical and practical growth,”⁴ Terry Cook warns that “a profession rooted in nineteenth-century positivism.... may now be adhering to concepts, and strategies and methodologies, that are no longer viable in a postmodern and computerized world.”

2. The Primary Objectives of the Study and the Research Questions

The primary objectives of the study were to assess the application of key components of archival theory to the management of sound records at Radio Botswana Sound archives

¹ UNESCO, “27 October Declared World Day for Audiovisual Heritage.” Also available at, http://portal.unesco.org/ci/en/ev.php-URL_ID=25525&URL_DO=DO_TOPIC&URL_SECTION=201.html

² Heather MacNeil, “Archival Theory and Practice: Between Two Paradigms,” *Archives and Social Studies: A Journal of Interdisciplinary Research* 1, no. 1 (September 2007): 517-545.

³ Terry Cook and Joan Schwartz, “Archives, Records and Power: From Postmodern Theory to Archival Performance,” *Archival Science* 2 (2002): 171.

⁴ *Ibid.*, p. 179.

2.1 Methodology

Data gathering methods of this study included interviews using face to face administered structured and unstructured interview questions, the researcher's personal observations and literature review.

2.2 Location of the Study

Radio Botswana Music Archives, Gaborone, Botswana

2.3 Population of the Study

All the three sound archivists who work for the Radio Botswana Music Archives were interviewed

2.4 Limitations of the Study

Bias is always a factor in both qualitative and quantitative research. I come from a mindset that believes that sound records are marginalized and have been given a raw deal in archives. The research could only get a snapshot of the Sound archives of Radio Botswana due to time constraints. The permission to undertake the research came very late. A more in-depth graduate work is needed to do justice to the study. However, despite these limitations I still believe that the findings of this study carry a high degree of validity.

2.5 Research Ethics

In most African government the media falls directly under the Office of the President. The Office of the President granted the research permit based on the topic. Consent was given based on voluntary participation.

3. Findings of the Study

3.1 The Acquisition of Sound Archives by Radio Botswana Archives

Respondents stated that they acquire their records from Live Special programs, for example when they cover the President opening a school or a hospital, pre-recorded material or in the case of music through donations from artists. Retired Broadcasting officers also freely surrender their records to the Radio archives.

3.2 Provenance, Respect des fonds and Original Order

3.2.1 Provenance

In archives the terms provenance, respect des fonds and original order form the core or nucleus of archival theory. Provenance refers to the agency creating the records. As the primary creator of sound records held by the archives Radio Botswana can claim the provenance of these records. However, in complex organizations, like Radio Botswana, with a lot of Directorates, one does not know whether to start with Radio Botswana as the provenance or to call a directorate provenance. 29 years ago Michel Duchein recognized this problem when he wrote that:

the agencies which possess numerous personnel and multiple powers are, in general, divided into areas called divisions, directorates, branches, sub-branches, departments, and so on, each of which exercises a definite part of the powers of the agency. It is clear that the records created by these divisions constitute organic wholes. There is, therefore, every interest in taking these divisions as the basis of internal arrangement of the fonds of the agency...⁵

This problem is bound to grow worse in the 21st century where multi-national chain store companies like Walmart, Coca Cola, or Apple have become dominant. So, in short, archivists now need to address the question of where provenance begins and where it ends in huge complex organizations that have huge record producing departments under them.

3.2.2 Respect des Fonds

One cannot discuss respect des fonds without reference to provenance. Duchein says that the majority of definitions of respect des fonds rests upon provenance to the point where countries of Germanic language refer to it as *provenienzprinzip* or *principe de provenance*.⁶ Mario Fenyo, archivist of America, dismissed the concept of a fonds saying that no one knows what the fonds means, not even the French who invented it,⁷ Fenyo's conclusion, however, has not stopped subsequent archival scholars from recognizing the fonds as one of the most important discovery in archival science. In fact, Duchein goes as far as saying that:

with few exceptions, the principle of respect des fonds is universally accepted as the basis of theoretical and practical archival science. Criticisms of the principle bear only.... on its applications and not on the principle itself. It is reasonable to think that it will never again be fundamentally questioned and that it constitutes a definitive fact of archival science.⁸

Heather MacNeil has also celebrated this principle. In "Archival Theory and Practice: Between Two Paradigms," MacNeil says, "the cardinal principle on which the medical profession is built is above all do no harm..... Similarly, the cardinal principle on which the archival profession is built is respect des fonds. While its proper application is frequently undermined by a seemingly endless list of realities- inadequate resources, authority, education, training- the principle is, in its own way, presents an equally worthy focus of archival inspiration."⁹

On this principle this research found out that twenty years ago with limited archival education and training, without even knowing that they were doing the right thing, archivists of Radio Botswana Sound archives practiced respect des fonds. They did this by grouping similar sound records together under the genres of Reggae, Rock and Roll, African music, etc, etc, This classification system led to the accumulation of records under this broad family names. All Bob Marley's records were kept together under Reggae. All the albums of the Beatles were together under Rock and Roll and African musicians music were archived under African jazz. This system made it easy to see similar records archived

⁵ Michel Duchein, "Theoretical Principles and Practical Problems of Respect des Fonds in Archival Science," *Archivaria* 16 (1983): 78.

⁶ Ibid., p. 73.

⁷ Ibid., p. 68

⁸ Ibid., p. 66.

⁹ MacNeil, "Archival Theory and Practice," p. 541.

together. This classification of the analogue sound records of radio Botswana affirmed the principle of “archives as universitatis rerum,” the indivisible and interrelated nature of archives, articulated by Luciana Duranti in her article on “archival science.”¹⁰ Respondents of this research lauded this system, saying that they never had any problem retrieving any song or tape that they wanted to play.

The success story that I have just described began to change drastically when management decided to replace these amateur archivists with graduates of library science. The principle of respect des fonds was replaced by chronological classification. All music albums entering the archives were no longer classified according to genre and artist but rather according to the date that the album entered the archives. Now retrieval and location of the song when it was needed became a big problem for the station.

This movement was a great regression that took Radio Botswana archives back to the days of Armand Camus and Pierre Daunou after the French revolution, when the duo decided to view records as discreet items and introduced chronological classification instead of records families. It took the courage of Natalis de Wailly¹¹ to reinstate the fonds as the most defining principle of archival science. In “Archival Science” Duranti talks of how ideas borrowed from the library methods created a real dichotomy between the theoretical and methodological concepts related to the nature, form, formation and management of archival documents against those that did not take into consideration the nature of archival materials as determined by the practical and administrative circumstances producing them.¹²

3.2.3 Original Order

The principle of respect des fonds is closely related to that of original order. Not only should records accumulate steadily under their families but the original order of their aggregation has to be respected as well. While this principle was found to be true for the analogue records of radio Botswana, it was not very for music that was archived digitally. Here, retrieval of any song posed no problem as long as the song’s name or artist was known. This finding is in line with Heather MacNeil’s assertion that in “paper based systems, where records are physically ordered in labeled files, usually in accordance with a classification scheme, the physical and contextual aspects of the records are intimately connected; original order has tended, for that reason, to be associated with physical arrangement. That association is no longer valid for most electronic records.”¹³

3.3 Appraisal

Rather than use the technical term “appraisal” respondents were asked if they ever destroy sound and if so under what circumstances. All the 3 respondents interviewed admitted that there are records that they erase a day or two after they been broadcast. What became clearer then in this research is that some programs are more important than others. This makes it possible to develop a comprehensive retention schedule for the sound records of Radio Botswana.

¹⁰ Luciana Duranti, “Archival Science,”

¹¹ The circular of 24 April 1841 signed by Duchatel, the minister at the Ministry of Interior codified the principle of respect des fonds.

¹² Duranti, “Archival Science,” p. 7.

¹³ MacNeil, “Archival Theory and Practice,” p. 526.

3.4 Arrangement and Description

The findings of this study showed that it was difficult to locate analogue records because of the system of chronological arrangement. There is a need to pay attention to the principle of respect des fonds. In the digital environment it did not matter whether records were arranged by the author or by the name of the creator because they would still be retrieved anywhere.

3.5 Access, Preservation and Digitization

Most archival repositories exist to provide access either to internal or external stakeholders or to both. Radio Botswana Sound Archives is no different. This research found out that it is easier for the staff to provide access to sound in the digital form than in analogue form for the simple reason that searching for the record in the digital environment takes split seconds while doing the same for analogue records might take forever. However, when asked about the sound quality and stability of the medium 2 out of 3 trusted the analogue records more. However, when asked whether they want to remain with analogue or digital records all the 3 archivists were eager to have their records digitized despite the fragility of the digital medium.

4. Conclusion

More detailed and sustained studies on Sound records and archival theory at a graduate are needed. However, preliminary findings of this study show that the basic tenets of archival theory are applicable to sound records in both analogue and digital formats. For me personally, these preliminary findings reflect that archival theory is relevant for all kinds of records, including sound records.

Editing Historical Music in the Age of Digitization

Jørgen Langdalen

The National Library of Norway

Abstract

Norwegian Musical Heritage—a program for preserving, editing and publishing historical music from Norway—benefits from digitization programs like that of The National Library of Norway. Digital media, in turn, opens new possibilities in the publication of historical music editions on the Internet. In both cases we see that digital technology draws new attention to the historical documents serving as sources of music edition. Although digitization strips them of their material form, they gain tremendously in virtual visibility. This visibility may contribute to a shift in critical edition, away from a predominant concern for the need to reach a single unequivocal reading, toward a greater interest for the real diversity of the sources. This broadened perspective may moderate the canonizing effect associated with traditional editions and open a more spacious field of interpretation and experience. In this way, more people may take part in the formation of cultural repertoires, in a media situation where such repertoires are no longer provided by traditional institutions, venues and media.

Author

Jørgen Langdalen is a senior researcher at The National Library of Norway, where he takes part in the preparation of the new Johan Svendsen Edition. Langdalen was trained as a music historian at The University of Oslo, where he earned his Ph.D. with a dissertation on Gluck and Rousseau.

1. Introduction

This paper presents a new initiative for preserving, editing and publishing historical music from Norway, called *Norwegian Musical Heritage*. The aim of the program is to make historical music available for performance and research in Norway and abroad.

I have been engaged by the National Library of Norway to take part in the preparation of a new, complete edition of the Norwegian composer Johan Svendsen's works. The Johan Svendsen Edition is a pilot project in *Norwegian Musical Heritage*.

Johan Svendsen made considerable success on the European music scene from the 1860s to the 1880s, both as a composer and as a conductor, and he was highly esteemed in America, too. In 1883, he accepted the position as musical director of The Royal Theatre in Copenhagen, and his career as a composer petered out, but his activity as a conductor on the international scene continued. Several prestigious institutions, including the Metropolitan opera in New York, tried to recruit him as a resident conductor, but he stayed in Copenhagen the rest of his life.

Svendsen left an impressive corpus of works in the nineteenth-century Romantic tradition—chamber music, symphonies, symphonic poems and concertos, and not to forget his masterful orchestrations of iconic pieces from the international Romantic repertoire.

My intention here is to show the influence of digital technologies and digital media on historical music edition. However, I will not only concentrate on the usefulness of digital technology in source studies and editing, but also emphasize the use of digital technology in the presentation of the editions on the Internet. Ultimately, I want to discuss the way digitization and digital media transform the meaning and role of historical music in present day music cultures.

2. Digitizing Programs

One important condition for the Norwegian Musical Heritage initiative is the digitizing program of The National Library of Norway. The library is a central partner in the initiative and holds the major part of the source material relevant for the editing projects.

The *primary sources* of any historical music edition include handwritten and printed scores left by the composer and his publishers. A manuscript in the composer's own hand is naturally an important primary source, but the production of printed scores, too, will normally be supervised by the composer and reflect her or his intentions. In twentieth-century music, even recordings are sometimes valuable primary sources, especially in the absence of scores.

Secondary sources include personal documentation such as correspondence and diaries, as well as documents from the publishers, concert institutions and so on, all of which may provide information about the genesis of the work. Printed matter such as newspaper reviews, articles and books may give supplementary information about the work's reception history, as do sometimes photographs and films.

The National Library of Norway has decided to digitize its entire collection and has come a long way in this enterprise. A major effort has been made in the field of musical manuscripts. Nearly 3,500 musical manuscripts have so far been digitized and published on the library's website, and many more are in the process.

Digitizing programs in libraries and collections across the world make modern music edition infinitely much more efficient than the old trade based on traveling and copying by hand. All kinds of source material used in critical edition are put at the disposal of the editors in digitized form.

Take Svendsen's symphonic poem *Zorahayda*, a piece of nineteenth-century musical orientalism, premiered in Christiania, Norway in October 1874.

3. Historical Music Edition and Canon Formation

In general, digitizing technology and digital media offer great new possibilities for preservation and remediation of historical documents. Yet, in the process, the cultural heritage itself is transformed. Any process of preservation and remediation of documents involves the intervention of new agents and interests. It is a highly selective process. The historical context that determined the meaning of the original documents—their status, accessibility, and use—is replaced by a context in which new interests define their meaning.

Thus, obviously, successful digital preservation of cultural heritage does not just depend on the permanence, authenticity, identity and integrity of the digital documents themselves, it also depends on their recontextualization in a new media situation—on the way they are put to use.

Historical music edition is a case in point. Traditionally, historical music edition has served as an authentication and authorization procedure aimed at preserving musical works from oblivion or from dissolving in an endless series of performances based on unauthorized scores. Although the sources—manuscripts and early prints—are the *sine qua non* of historical edition, it is in the nature of things that they are subordinate to the finished, edited version. The final edition, as soon as it has been completed and published, *replaces* any source. What is then presumably revealed, is the work “itself”, considered as a constant and invariable entity buried under the pile of sources. The “work” in this sense is, of course, an illusion.

The establishment of the text of a musical work—its “wording”, so to speak—is the central task in editing. The critical apparatus of the edition—in the name of source criticism and sound philological craftsmanship—will describe and discuss any possible reading and leave no details in the source material uncommented. Yet, many conflicts in the sources only find their resolution through an interpretive effort by the editor. Rather than openly discussing this interpretive effort, traditional editions refer to the authority of the underlying mysterious *Urtext*. The overwhelming yet deceptive impression left by multi-volume, monumental collected editions is that the editors have reached the ultimate and indeed only possible reconstruction of the opus.

There are good practical reasons to reduce the number of possible readings and present the performers with an as unequivocal score as possible. However, the quest for the ultimate reading goes far beyond this practical concern and into the metaphysics of the musical work.

By progressively relieving historical works from their sources and gradually uncovering the underlying “work”, historical editions contribute to the establishment of a *musical canon*. This canonizing effect—anticipated in the moment a decision is made to edit the work of a certain composer and not somebody else, and to edit certain works within the corpus of this composer and not others—can be avoided in a new approach to the sources, in the age of digitization and digital media.

In the case of Johan Svendsen, any version of his collected works will have to relate to the question of a Norwegian canon. The Svendsen edition could be the first step in a long awaited revaluation of the role of music in what is often referred to as the Norwegian Golden Age—the age of Ibsen and Grieg. Svendsen’s position in this context is somewhat ambiguous and his contributions to this heritage not easily subsumed under the usual banners.

The new edition will include music that do not necessarily support the prevailing picture of Svendsen as a representative of National Romanticism, such as his unpublished dances and marches for orchestra, written for the mid-nineteenth-century entertainment scene in Christiania, the celebratory cantatas and marches commissioned for political events in Norway and Denmark.

4. The Return to the sources in Music Edition

Source-based research in general can be an effective means to recontextualize historical art encumbered by ideologies of later times. Although critical edition—the source-based discipline *per se*—is inevitably involved in canon construction, it may also have the power to change our view of history.

In this, it would answer a call in present day scholarship for a new return to the sources and new attention to the materiality of the cultural heritage.

Obviously, the renewed interest for sources and materiality in the humanities finds an ambiguous allied in digital technology. Digitization speeds up the publication and dissemination of historical documents, but in the process they are stripped of their physical and material form, their *historical* form. The specific nature of traditional materials such as parchment, paper, vinyl, magnetic tape and so on is lost, as is the physical, corporeal quality of specific media such as the book, the map, the record and the musical score. But the notion that remediation does not alter or influence the message, and that the message remains the same no matter which material form it assumes, seems very deep-rooted.

The materiality of the cultural heritage has become the object of renewed interest in critical thought and historical research. If the digital media of the 21st century and their power to transform practices and experiences attract scholarly interest, so do the analogue media of previous centuries. There is a growing awareness that historical documents and artefacts do not only communicate a coded message but also

possess a material presence. To read a historical document on the screen may give you the content of its message, but to hold it between your hands is to face the unknown or forgotten.

Book history is one of several new fields of research in which the meaning of materiality is explored. The methods of book history draw on a variety of existing disciplines such as media history, classical bibliography and, indeed, edition philology. Book history explores the physical properties of the text in order to reveal what it says, apart from what is written in it. The binding, paper quality, typography, the illustrations, the price, the subscription list, the number printed—these are all factors that add to the meaning of the book by showing glimpses of its use, distribution, readership and status. In this way, books are made to tell stories about power, politics, economy and technology—and about marginalization and canonization.

The return to the sources seems to be felt more urgently precisely in the age of digitization. What is sought for is the new sensibility in the encounters with the source material in their concrete materiality. In the field of music, the need to return to the sources is even more urgent, since this particular field has been so thoroughly digitized.

Although the massive distribution of digitized documentation on the Internet poses some paradoxes for source-based studies, I choose here to emphasize the positive side. And although even the best scan will not be able to reveal all the secrets of the physical document, the sources sometimes seem to regain some of their mysterious presence in present days digital media.

5. Digitization and New Digital Media in Music Edition

The return to the sources of historical music is increasingly made possible by digitizing programs in the institutions that hold the material. Thanks to digitization and digital media, editors and researchers can make use of digital documentation—especially in the field of secondary sources—that in earlier times would not have come to light, or been obtained only through very time-consuming archival studies.

However, the full power of digitization and digital media lies in their potential to present the products of historical edition to the public in new and innovative ways.

On a basic level, digital web-based editions can make use of multimedia technologies that give the public a deeper insight into all details of the source material, but without renouncing the clear and playable score. Any amount of supplementary information can be made directly accessible in the text itself through pop-ups and hyperlinks. Any number of primary sources can be made available *in extenso*, in digitized form on the website, and be linked to the new edition in many ways. Different versions of the finished score can be presented to performers, and engage them in an interactive use of the edition.

Digital technologies and media offer even more possibilities in the field of secondary sources. Facsimiles of handwritten and printed material—letters, diaries, newspaper articles etc.—as well as digitized photographs, films and other documentation may be able to present a totally new picture of the works in question, and tell a new story about the genesis and history of the work. The documentation material can be arranged in interactive patterns that do not necessarily support a single interpretive perspective but allow the user to find her own path and pursue her own line of interpretation.

In this way, the historical documents will have fulfilled a quite different purpose than the occasional facsimile reproductions known from traditional editions.

Although historical edition is inevitably a selective process guided by interest, the canonizing effect can be controlled by employing resources that emerge from the digitizing process itself. Digital music edition, presenting the edited work in digital form on the Internet, may feature any number of digitized

sources and a wealth of other digital resources that provide documentation for the creation of the work as well as its history of performance, publication and reception. In this way, digital editions give new possibilities for studying the processes of canonization and marginalization themselves.

6. Historical Music in the Digital Public Sphere

The new potential to recontextualize historical documents is the result of a media situation in which information and communication, experiences and entertainment circulate in new ways. New digital media bring about changes of perspective and altered sensibilities. They create a public sphere in which traditional mechanisms for establishing musical repertoires are transformed or replaced. The institutions, venues and media that have traditionally made up the public sphere have not been able to sustain their traditional roles and positions. The conditions for the formation of historical knowledge are changing, and new mechanisms for canonization and marginalization arise.

New digital media have brought about a revolution in production and distribution of music, including historical music of all kinds. The prominent role of music in contemporary media culture—digital dissemination of sound and images and other forms of documentation, including sheet music—represents a new and more unpredictable context for editing historical music. In the time to come, any ambition to present authorized versions of historical works and thus to define a canon, seems to be limited.

Digitization and digital media bring about fragmentation but also democratization. This ambivalent situation involves a particular responsibility for researchers and publishers. As historians, we have an obligation to not just preserve the permanence, authenticity, identity and integrity of the historical documentation that we present, but to open a space for their interpretation.

Developing and Implementing a Digital Video Repository for Legacy Dance Documentation

Dance Heritage Coalition's Secure Media Network

Lauren Sorensen¹ and Tanisha Jones²

¹Bay Area Video Coalition, USA; ²Jerome Robbins Archive of the Recorded Moving Image, The New York Public Library for the Performing Arts, USA

Abstract

The Dance Heritage Coalition and Bay Area Video Coalition are collaborating on a Mellon Foundation funded project to digitize, preserve, and make accessible analogue and digital tape-based materials of key importance to dance heritage. In creating a digital repository, materials are managed using a suite of tools and standards, including OAIS standard (ISO 14721:2003); the project currently involves moving from Fedora to Archivematica repository software, and custom software tools for file analysis and digital collection management. The goal of this paper is to outline the technological and workflow related challenges faced in implementing and building the repository, focusing on those related to audio-visual preservation, and specifically addressing challenges in building an archival workflow without documented standards of practice for digitization of analogue and digital videotape in the moving image archiving and preservation field.

Authors

Lauren Sorensen is Preservation Specialist at Bay Area Video Coalition, and co-chair of the Association of Moving Image Archivists' Independent Media Committee. Previously she has held positions as Assistant Director at 16mm film distributor Canyon Cinema, research assistant for Preserving Digital Public Television, an initiative of NDIIPP (National Digital Information Infrastructure and Preservation Program), and has acted as preservation consultant for a number of independent film and video collections. Lauren holds an M.A. from New York University in Moving Image Archiving and Preservation. In her current role at BAVC, Lauren manages digitization, technical development, and documentation of the Secure Media Network.

Tanisha Jones serves as Director of the Jerome Robbins Archive of the Recorded Moving Image in the Jerome Robbins Dance Division at The New York Public Library for the Performing Arts, and has worked in the Dance Division since 2007. As of 2009, Ms. Jones is Consultant to the national non-profit alliance organization Dance Heritage Coalition's Secure Media Network project, which aims to provide secured access of digitized archival dance materials from various national dance collections, archives, libraries, educational and research centers. Additionally, Ms. Jones has presented her work at professional conferences and organizations such as the American Library Association (ALA), Association of Moving Image Archivists (AMIA), Joint Technical Symposium (JTS), and the Archivists Round Table of Metropolitan New York (ART). She served on the Board of Directors of the New York-based nonprofit organization Independent Media Arts Preservation, Inc. (IMAP) from 2005 – 2009, and since 2009 is on IMAP's Advisory Council. In 2011, Ms. Jones was an adjunct professor in Tisch School of the Arts' Moving Image Archiving and Preservation (MIAP) program at New York University; the program from which in 2005 she received her M.A. as a graduate of the MIAP's inaugural class.

1. Introduction

The Dance Heritage Coalition and Bay Area Video Coalition are collaborating on an Andrew W. Mellon Foundation funded project to digitize, preserve, and make accessible analogue and digital tape-based materials of key importance to dance heritage. In creating a digital repository, materials are managed using a suite of tools and standards, including OAIS standard (ISO 14721:2003) based systems FEDORA and now moving to Archivematica repository software, and in navigating the transition between a Fedora system, using custom software tools for file analysis and digital forensics specific to moving image video files.

Dance Heritage Coalition is the sole national non-profit alliance of institutions holding significant collections of materials documenting the history of dance. Its mission is to preserve, make accessible, enhance and augment the materials that document the artistic accomplishments in dance of the past, present, and future. The DHC was founded in 1992 to address problems in documenting dance and preserving the record, problems which were identified in a study commissioned by The Andrew W. Mellon Foundation and the National Endowment for the Arts. This study, titled *Images of American Dance*, recommended the formation of an alliance of the nation's major dance collections to facilitate communication; develop national standards, policies and priorities; and implement collaborative activities and projects in the fields of dance preservation, documentation, and access. In keeping with this history and mission, in 2009 the DHC embarked on an ambitious project to preserve and make accessible vital dance documentation, rare performances, and other notable works via a digital preservation repository of moving image materials as well as a corresponding web-accessible repository available on-site at DHC member archive institutions.

2. The Problem: Deteriorating Carriers for Moving Image Dance Documentation

Analogue videotape is a commercial storage medium initially used in television production and broadcast contexts starting in the 1950s with 2" Quadraplex videotape. In subsequent years, videotape in other forms became widely used as an affordable format for artists, and quickly shifted to consumer applications as well. Performance documentation starting in the late 1960s was documented on the Sony Portapak, one of these affordable means to record video. Subsequently, for the dance community, many dance companies, dancers, dance studios and festivals used this affordable, consumer videotape over the course of many years; the materials within DHC member archive collections, according to the Secure Media Network union catalog (publically available at <http://archive.danceheritage.org/>) are in the range of mid-1960s to the present. While analogue videotape continued to become less and less expensive, the formats developed were never consistent; they relied on a company or commercial entity to maintain and manufacture equipment, and as soon as new developments in technology occurred, a new format was born replacing the old which was then no longer supported by the company producing the videotape and videotape playback machines, which created opportunities in quality and technological advancement but challenges for sustainability. Non-profit organizations such as dance institutions saw opportunities to document practices, rehearsals, dancers in these formats and so subsequently, looking at these collections, archivists and subject specialists can see that this practice was incredibly valuable to the history of dance. However these carriers of vital documentation and history face a lifespan of just 15-20 years under *ideal* temperature and humidity, a practice which is oftentimes outside the purview of non-profit organizations, and so as a result these valuable assets become a high priority for preservation purposes.

The Dance Heritage Coalition identified this as an issue concerning dance legacy: analogue video assets are the most at-risk, endangered by obsolescing equipment, physical deterioration, and their key importance to ; a three-fold relevance to what the project wants to support, and the DHC board chose BAVC as a partner because of complementary interests in pursuing open-source standards and creating a community around preservation of analogue video for under-served organizations like non-profits and artists. BAVC has developed a standard and accepted workflow for preservation of analogue media assets that would be adopted by Dance Heritage Coalition digitization hubs around the country, giving the organization the opportunity to pursue preservation without the need for a vendor in all cases. Preservationists determined that the most at-risk videotapes would be handled in-house by BAVC, as many times they require expertise that is not transferable to digitization fellows who would be handling digitization of materials that are less susceptible to factors such as sticky shed syndrome, and particularly fragile equipment.

The project also incorporated choosing and developing a plan for using repository software to manage the assets; up to now, we have been using a system of command line software tools to ensure integrity of the files and our own initially FEDORA-based, documented system for creating Submission Information Packages.¹ A project goal was to make our SIPs as robust as possible given that the system is being “incubated” and developed at BAVC, and that this system will move to a DHC partner organization, perhaps with a different infrastructure, it was important to employ a well-documented and flexible Open Archival Information System² compliant system that would then be easily navigable when the system moves to another organization.

Since many dance companies, festivals and other organizations and artists producing videotape content used consumer videotape formats to document performances, rehearsals and other important dance related documentation, many of these materials have ended up in cultural heritage institutions charged with preservation of artists’ records, dance companies archival materials donated include such materials and so on. However, many of these institutions, while charged with the preservation of these videotape formats do not have the facilities, technical knowledge base and resources from which to actively preserve these important works; Dance Heritage Coalition identified this issue in many of their constituent organizations’ collections and took up the challenge to draw resources around this issue and pull technology and archive partners together to come up with a solution, and develop and build a digital repository for digitized moving image materials comprised of dance documentation from partner archive organizations.

3. How can a non-profit organization establish and sustain a digital repository?

The goal of this paper is to give a case study and outline several challenges in the project, which is currently in the late discovery phase. During the course of this paper we will go into detail in these areas of the project’s development: *Developing a Union Catalog*, *Metadata Standards*, *Workflow Development*, *Digital Preservation of Moving Images* and *Building Geographically Diverse Digitization Hubs & Digitization*.

¹ “Reference Model for an Open Archival Information System.” The Consultive Committee for Space Data Systems. CCSDS 650.0-M-2 <http://public.ccsds.org/publications/archive/650x0m2.pdf> (Accessed 08-01-2012)

² Ibid.

4. Developing a Union Catalog, Metadata Standards, Workflow Development

The first goal of the project was to gather catalog records and build a catalog resource for researchers and archives' access online and for access to the DIPs (Distribution Information Packages). Next, we refined and documented BAVC's own analogue videotape preservation workflow to plan for and develop workflow for digitization hubs in DHC partner institutions, and find a way to support that infrastructure. Developers on the project reviewed the goals and developed a collecting policy which outlines the content and factors around choosing elements for preservation and named constituents and user profiles such as submitting archives, users and so on. This gave the project a basis for reviewing and then developing technical goals for the SMN that ended up speaking directly to the collecting policy.

The digital age gives cultural heritage institutions and institutions options opportunities but also so many choices that, in order to become meaningful and sustainable long-term must be strategically made. The SMN project has functioned with this in mind—building out a system where the goal is to be flexible and sufficiently documented to be sustainable moving forward. The project took on the standard OAIS structure (Open Archival Information System model) standard for digital archives as a basis for developing the project, see Figure 1 for a visual model of the system. The system is currently transitioning from FEDORA to Archivematica system.

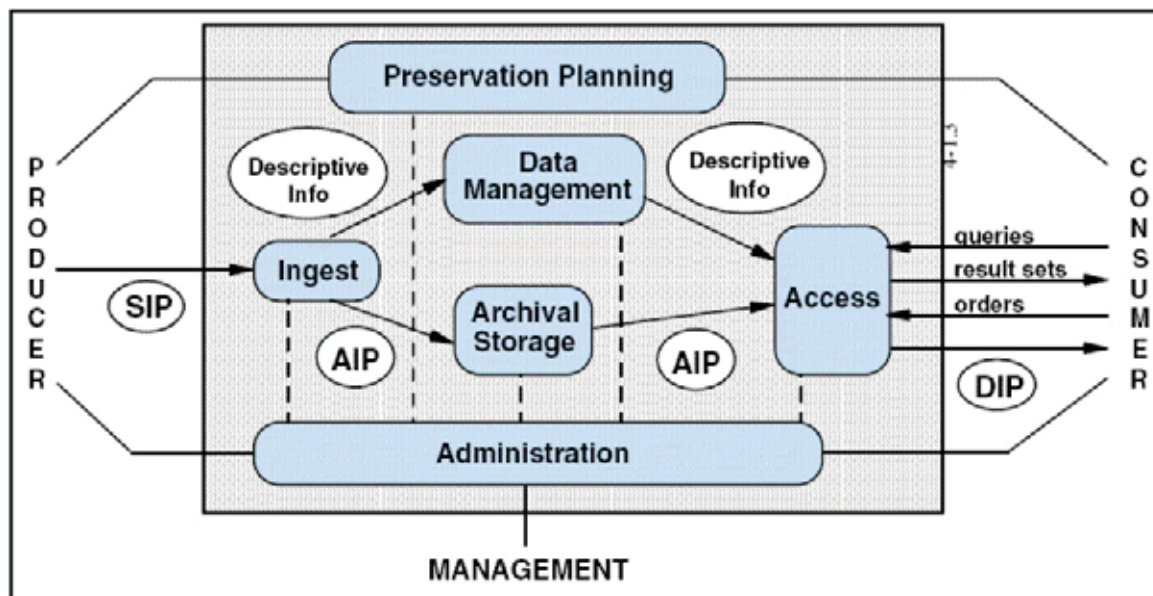


Figure 1. Open Archival Information System (OAIS) Model.

During this development and discovery phase, the SMN access and preservation repositories are located at Bay Area Video Coalition, a 35 year old non-profit based in San Francisco known for its work as a non-profit center for analogue video preservation; the digital repository development component and metadata maintenance would be managed by David Rice, consultant and technologist currently an Archivist for City University of New York, formerly of AudioVisual Preservation Solutions and WNET.

The project began with many of the usual issues seen by not-for-profit organizations in the United States: limited funding, organizational infrastructure that is largely reliant on grants and a shoe-string budget; a staff of one whose main job description involves grant writing and project management. So the

collecting policy and plan that was made had to include partner archives' active involvement and real collaboration as an alliance of archives. One of the questions early on involved obtaining item-level metadata records from submitting institutions' archives own cataloging efforts. This would then allow access for researchers and provide a central location where archivists from existing institutions could access records from partner organizations and researchers could see where dance documentation double as a place where Distribution Information Packages, in the form of streaming video files would live and be accessible via research request practices at individual DHC member organizations. The structure was well suited because for the project would include different instantiations of one asset; a feature supported by Public Broadcasting Core, the schema that was used as the basis for this web-based application. Our hope was that FEDORA could operate as a repository software tool for the project, however the large file sizes required by the project were not workable with FEDORA; this led us to talking with colleagues and approached Archivematica as a possible solution for our large file size maintenance. Currently we are developing a plan to transfer our current FEDORA based SIP structure to Archivematica SIP structure; learning in the process how important the skillset of working in a command line environment can be for a digital preservationist or collection manager. With a great support for file fixity checks as well as an easy-to-use interface, as well as collegial experience using the Archivematica suite of tools with larger file formats, we feel confident that this move will be a positive addition to project. In the past, working with FEDORA, the repository software was only able to "point to" the files instead of actively manage them; unfortunately a ticket put in to address this issue of problems related to large files has not been addressed since 2009.³ Our own testing found that the upload progress lagged and never completed upon upload of large (50 gigabytes or more) files.

5. Digital Preservation of Moving Images

Digital preservation in general is a challenging topic because of a variety of different factors, including openness, sustainability; but digital moving image preservation creates a number of even more challenging variables, including a consistent file format choice for analogue-to-digital preservation; large, high quality file management by institutions that may have technology personnel accustomed to smaller files managed on servers, such as documents or compressed video; lack of standardization in the moving image preservation field, fixity issues made more challenging by large file size; storage capacity and relationship between the need for preservation and not-for-profit budgets; infrastructure and administration and complex copyright challenges.

Digital preservation of moving images is a facet of a larger field, and this project gave BAVC and DHC the opportunity to approach challenges related to developing procedures and protocol around file-based digital moving image preservation. In the course of pursuing analogue video preservation, we learn that in this context, traditional archival principles had to shift, especially in terms of authenticity; keeping an artefact its original form essentially relegates the content contained therein to the wastebasket because of analogue media's unstable carrier. Kenneth Thibodeau introduces his article for CLIR, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," by using a decidedly analogue example: the periodic table; this example is meaningful not only by the content but by the structure; much like moving images which are not only visual images but visual images with a specific frame rate, in a certain colour space and so on. Because of the technological challenges we may

³ FEDORA Commons Wiki"2009-06-09 - Special Topic - Large Datastreams."
<https://wiki.duraspace.org/display/FCREPO/2009-06-09+-+Special+Topic+-+Large+Datastreams>.

not see that we can preserve not necessarily the exact form of that periodic table but the characteristics of it that make it meaningful. He states:

The periodic table was created a century before computers, and it has survived very well in analog form. Thus we cannot say without qualification that the variety and complexity of digital objects must always be preserved. In cases such as that of the periodic table, it is the essential character of the information object, not the way it happens to be encoded digitally, that must be preserved. For objects such as the periodic table, one essential characteristic is the arrangement of the content in a 2-by-2 grid. As long as we preserve that structure, we can use a variety of digital fonts and type sizes, or no fonts at all—as in the case of ASCII or a page-image format.⁴

So how do we go about determining what is the essential character of a digital object, or its essence from the analogue source (or in the case of DV, the data contained on the tape carrier), and therefore what to preserve? And then, what do tools are needed to preserve that essence?

6. What is the essential character of the object being preserved that the preservation system, digitizing technicians, and workflow will need to support?

After need was identified, BAVC embarked on a study published by the American Conservation Institute's Electronic Media Group that examined what codecs are appropriate for the preservation of dance documentation from an analogue source. The goal of the paper was choosing a best format for the project, but also act as a case study. What codec could work the best for a preservation repository of this kind, and what characteristics of the original analogue source would be preserved with accuracy?

The Secure Media Network requires archival quality digital files to be stored in a digital repository well as streaming video to be viewed—because of copyright restraints—within the walls of the member organizations. In the spring of 2009, in preparation for this undertaking, BAVC performed this study to evaluate current commonly used codecs for the reformatting of analogue video videotape. Nine participants were selected for this study. The participants included video professionals, colorists, archivists, and a dancer. In addition to informing our decisions regarding codecs for the Secure Media Network, BAVC hopes that this study will provide an additional data point and case study for archives and other organizations to consider when planning their inevitable migration to digital files. Each participant spent about two hours comparing video clips. The participants were asked to note the following:

1. Visual difference between clips. Was any difference noticed?
2. Compression artefacts observed. Compression artefacts include blockiness, blurring, etc. Because these codecs are considered to be production quality, these artefacts were expected to be minimal.
3. Chroma reproduction. Were there any shifts in hue or chroma level from one clip to the next?
4. Luminance reproduction. Were there any differences in luminance from one clip to the next?
5. Any other artefacts. Were other differences noticed?

⁴ Thibodeau, Kenneth. "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," CLIR and the Library of Congress, *The State of Digital preservation: An International Perspective*, April 2002.

Answers to these questions were noted by the operator and are included in the report published in the Electronic Media Review, Vol. 1. Overall, most participants noticed major differences only in the H.264 files and the IMX files. Some participants, even between 8-bit and 10-bit uncompressed files, noticed minor colour differences but all stressed that it was very difficult to see these differences and were very subtle. Time base correctors, playback equipment or monitor issues can also introduce these types of colour changes. In general, artefacts such as blockiness, blur or motion artefacts were not seen when comparing 8-bit and 10-bit uncompressed files, but were seen in H.264 files. The artefacts that were seen were subtle “softening” in varying degrees on certain clips and a reduction in apparent noise, which led some participants to conclude that the compressed file was “better.” Since lower quality analogue sources were used, most of these tape formats exhibit signal problems which can be difficult to distinguish from other artefacts that may be introduced by digital compression. Based on this study, BAVC recommends 10-bit uncompressed files for any visually significant recording; the case for performing arts documentation. These lossless high quality files are also useful in any future transcoding that may need to occur for preservation purposes of the digital files; lossy codecs, when transcoded lose information and result in poorer image quality and loss of information moving forward.

10-bit uncompressed files coming from an analogue source, must fit within a certain video bandwidth so to align these results with we have to purchase capture equipment that will align with these requirements and choose a format to digitize and capture to that will adequately represent the picture. A chart developed for the Conservation Online web site, maintained by American Institute of Conservation’s Electronic Media Group was useful tool in evaluating this along with capture card specifications.⁵

Specifications for the files formats from an NTSC analogue source can be found below:

Video essence

Audio essence

File extension: *.mov or *.avi			
Format (wrapper) profiles: Quicktime or Audio / Video Interleave			
Codec & Codec ID	YUV v210, AJA Video Systems Xena	Codec & codec ID	PCM
Bit rate mode	Constant	Bit rate mode	Constant
Bit rate	224 Mbps (approximate)	Bit rate	2300 Kbps (approximate)
Original width	720 pixels	Channel(s)	Up to 4 channels
Original height	486 pixels	Sampling rate	48.0 KHz
Display aspect ratio	4:3	Bit depth	24 bits
Frame rate	29.970 fps		
Standard	NTSC		
Colour space	YUV		
Chroma subsampling	4:2:2		
Bit depth	10		

The metadata relating to the digital object, once processed for ingest includes metadata from Mediainfo, an open-source tool that also allows for checking of file characteristics, like those above. PREMIS is also

⁵Conservation Online. “Range of Video Formats,” http://videopreservation.conservations-us.org/dig_mig/video_formats_v4_850.html (Accessed 2012-07-08).

mapped from Mediainfo, so this tool is used for multiple purposes: digitization fellows are trained to review this information to confirm aspects of the file requirements are in line with what was produced. It is also mapped using command line tools to PREMIS, a metadata standard used in many digital repositories.

Our project also utilized a frame-by-frame checksum reporting command line tool which allows users to check frame by frame checksums which is helpful for fixity purposes through the software application ffmpeg. What are checksums? David Rice, our project technologist and metadata specialist wrote an article in a recent IASA (International Association of Sound and Audiovisual Archives) journal titled “Reconsidering the Checksum for Audiovisual Preservation: Detecting digital change in audiovisual data with decoders and checksums” and provides this outline:

A checksum is small data value computed from a given amount of data, such as a file or bitstream, for the purpose of facilitating the future ability to detect changes in that given data. The generation and verification of checksums for digital archival holdings is a central principle of digital preservation and enable archivists to trust that data held within an archive is the same data that was received by the archive. Although checksum wrangling is typically a behind-the-scenes process within digital storage systems and repositories, these values are worth a closer look. The checksum value is generally expressed in hexadecimal representation (aka base 16) comprised of the numbers 0 through 9 and the letters A through F... [If a] file changed, whether through manipulation, bit rot, or data corruption, then further evaluations of the file would produce a different checksum value. The mismatch of a newly calculated checksum and a stored checksum produced earlier is an alert that data under care has been changed.⁶

Technicians create an MD5 checksum upon completion of digitization; this is reviewed following transmission of the SIP to the collection manager, and Archivematica uses MD5 checksums as well. Because the files that result are so large (an aspect which will be reviewed later in this paper) submitting a SIP over a network is not possible for many non-profit organizations. So we submit SIPs by mailing hard drives from the digitization hub locations to BAVC, where the repository lives; please see Figure 2 for a picture of the SMN workflow. Additionally, once a SIP is ingested into the repository, an application called rsync is used to copy the files from the on-site server controlled RAID to an off-site data center which houses another server controlled RAID (formatted as RAID 5). Rsync then generates reports that the collection manager may review for digital object maintenance outside of the repository structure in order to determine if any changes to the digital objects have occurred and take steps to repair or address the issue from there.

The access repository is separate, but managed in the same suite of scripts that create a fully realized SIP following digitization by hub technicians. It is the same interface as the union catalog, a PBCore database that was originally developed by Roasted Vermicelli LLC for the WNET Digital Archive. It was further developed in 2009–2011 for the Dance Heritage Coalition Secure Media Network and graphic design was handled by Blue Griffin Design Company. In keeping with the not-for-profit institutional context for this repository project, the program is free software: it can be redistributed and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or any later version, and is held under a GNU license. Code

⁶ Rice, David. “Reconsidering the Checksum for Audiovisual Preservation: Detecting digital change in audiovisual data with decoders and checksums” IASA Journal 39 (2012).

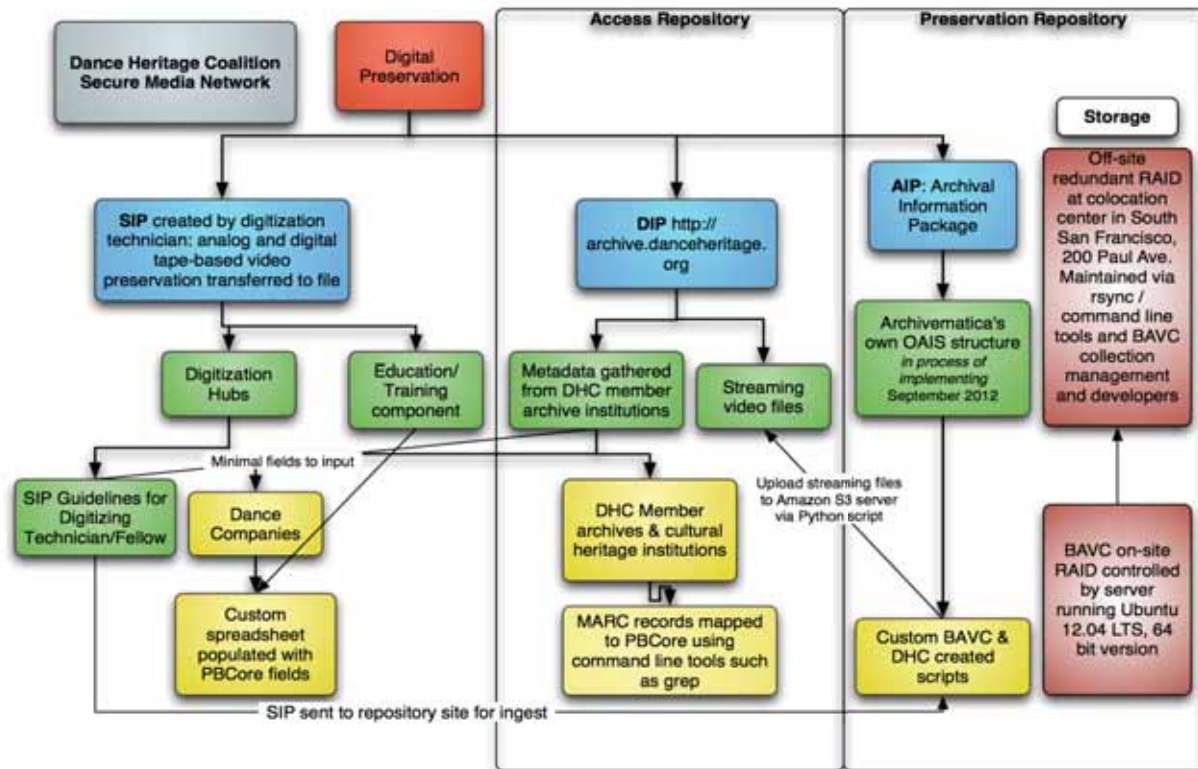


Figure 2. Digital Repository Workflow for DHC Secure Media Network.

in the vendor and gems directories of this project was created by other authors under potentially different Free Software licenses. This work employs PBCore. The PBCore (Public Broadcasting Metadata Dictionary) was created by the public broadcasting community in the United States of America for use by public broadcasters and others. Initial development funding for PBCore was provided by the Corporation for Public Broadcasting. The PBCore is built on the foundation of the Dublin Core (ISO 15836), an international standard for resource discovery, and has been reviewed by the Dublin Core Metadata Initiative Usage Board. Copyright: 2005, Corporation for Public Broadcasting. This software is powered by Ruby on Rails, Apache Solr, MySQL, nginx, and other third-party modules.

This application works well for inputting metadata because of the web-based interface. Protocols for which fields were required are outlined in the project's SIP (Submission Information Package) guidelines, a resource and training tool for digitization fellows who preserve materials at member archive hubs (see Figure 2 for a visual on where these hubs fit into the over-arching workflow of the project). However, since the project has developed, the application is no longer supported and features are no longer being handled by developers, so the SMN will likely shift to work within another CMS, such as Drupal with modules that support XML and streaming video. We are hoping for another form-based user interface for metadata input and mapping which will continue to allow ease of use for our digitization fellows to handle. There is a newer version of PBCore standard (2.0) which developers on the project are working on mapping from PBCore 1.3.

7. Building Geographically Diverse Digitization Hubs & Digitization

Next the project focused on implementation of digitization “hubs” or small centers within DHC member organizations around the country where digitization fellows with either a background in preservation and archiving or video production would create digital objects from an analogue source, preservation and technical metadata and supply Submission Information Packages for the SMN digital repository. These hubs are largely small workstations within non-for-profit DHC member archives, and BAVC preservationists as well as on-site staff provide supervision and training; fellowships are often helpful in growing skillsets for these archivists but also in providing feedback on the project more generally; the goals of the project grew to incorporate an education component that has proved vital to the project’s development on the whole.

Submission Information Package guidelines were initially based on FEDORA requirements and include detailed guidelines around preservation of analogue video; an at times slow process which involves identifying risk factors and real-time supervised transfer; because these materials have an estimated lifespan of less than 15 years, the risk for deterioration and aging playback machines, BAVC additionally is providing the digitization end of the repository project with general guidelines, protocol and contacts for engineers around the country who would be perhaps able to repair and maintain the obsolete equipment in use by digitization hubs. These engineers are becoming more and more difficult to locate and the hope is that through development of this project and adding new digitization hubs we will be able to make contact with more experts in this field and glean important knowledge while putting them to work in maintaining this aging equipment.⁷ The guidelines also provide information on identifying formats, inspecting analogue videotape, terminology and expected outcomes in terms of digitization package creation for submission to the SMN. Since the SMN is going through a transitional phase for our preservation repository from FEDORA based system to one built on Archivematica, we are now updating the guidelines to reflect this shift. Additionally, dance companies and studios’ with “default” archives are submitting tapes for the SMN archive, and so much of this material will be handled via simple spreadsheet input and the fellows will then create new records within the PBCore application—for the purposes of the access repository but also for identifying the asset to be digitized and starting out the workflow in a consistent way.

For our second digitization hub, at the DHC member archive, the Dance Notation Bureau in Manhattan, New York the project took a slight shift; digital videotape was introduced and a Linux-based system was developed and a preservation fellow was trained. DV, unlike analogue video is already comprised of data, bits, only on tape form. David Rice while at consulting firm Audiovisual Preservation Solutions, published a white paper⁸ referring to DV as an at risk medium, and identifying the need to transfer data off of the tape; he then developed a tool called dvanalyzer; this tool was implemented as a quality control tool and incorporated into our technician guidelines for the project.

Incorporating DV format, we saw this as an important step for diversifying the media held by the repository. While there are very specific standards for analogue sources, we want to make sure any digital

⁷ As a skillset, videotape playback equipment repair is becoming fast as obsolete as the equipment, in fact some archivists and preservation activists have worked on projects to document and record oral histories with these experts – some of these resources can be found through the Experimental Television Center’s web site Video History Project (<http://www.experimentaltvcenter.org/history>).

⁸ Rice, David and Chris Lacinak. “Digital Tape Preservation Strategy: Data or Video?” <http://www.avpreserve.com/dvanalyzer/dv-preservation-data-or-video/> (Accessed 2012-08-02).

source ingested is evaluated on its own characteristics; DV for example, is an open standard codec and even though it is lossy, it is widely adopted and any transcoding to our analogue standard would result in “empty bits” or a bit depth that is higher than what is represented visually, and perhaps an inauthentic colour space. More information on this topic can be obtained from “Digital Tape Preservation Strategy: Data or Video?”,⁹ and working with Archivematica allows us the flexibility to work with different types of preservation standards.

SIP packages from the digitization hub centers are then sent via hard drive overnight service or 2-day to the repository center, currently at BAVC. This is where custom scripts are run on these SIPs to prepare them for ingest; moving forward the intention is to train staff at digitization hubs to develop skillsets to implement this preparation work moving forward, however this will likely have to happen after Archivematica 1.0 is fully implemented. Upon ingest, following checksum review, the SIPs are copied to a server running Ubuntu Linux 12.04 LTS and then the rsync command is used in concert with the cron command and various specific specifications to copy the SIPs nightly to the off-site server storage at a data center in South San Francisco. Due to the large size of the files we are working with, it was important to choose an off-site data center where BAVC’s city-implemented fiber connection could be utilized. This aspect will have to be re-evaluated once the project incubation period is over, as fiber is not a common nor cost-sensible solution for many non-profit institutions.

8. Conclusion and Next Steps

The SMN is a project that has a lot of potential moving forward for the dance community and highlighting dance heritage within the digital archives community. The next steps involved in the process have to do with implementing many of the recommendations related to the TRAC (Trustworthy Digital Repositories Audit & Certification) checklist. From a technological standpoint, the checklist remains useful in providing the project with context and recommendations for a framework moving forward; another goal is evaluating and re-evaluating financial sustainability, support for a large technological infrastructure, and continuing the initial goal that set the project to start, namely preserving deteriorating and obsolescing audiovisual media. The OAIS standard was also recently updated, and once our repository software is implemented fully, the project plan is to document using a Use Case model every step of the standard as another auditing procedure and way to document for future caretakers of the collection; Archivematica has done a version of this Use Case, however since our model includes several custom, audiovisual specific workflow steps, we intend to amend it.

Moving forward, Archivematica 1.0 will be implemented early next year, having created from their systems a new way of forming our ingest SIP packages; and creating our custom scripts, some altered from some Archivematica micro-services detailed in Figure 2, which gives a larger picture of our overall transition from FEDORA to Archivematica. Having met with colleagues at the City of Vancouver archives who work with losslessly compressed video in the context of Archivematica, our hope is that the microservices chosen and the option for normalization will give us sustainable and flexible options moving forward in working within the Archivematica system. We also hope that this project results in more traction and conversation around pursuing preservation of analogue videotape assets in archives, and what digital preservation can mean for historical documentation on these rapidly decaying carriers of moving image heritage.

⁹ Ibid.

Beyond Access:
Digitization to Preserve Culture

Safeguarding of the Portuguese Language Documentary Heritage

The Lusophone Digital Library

Fernanda Maria Melo Alves,¹ José António Moreiro González² and José Manuel Matias³

¹Department of Information Studies. Carlos III University of Madrid, Spain, fmelo2@hotmail.com

²Department of Information Studies. Carlos III University of Madrid, Spain, jamore@bib.uc3m.es

³International Association of the Lusophone Digital Library, Lisbon, Portugal, jlmatias@netcabo.pt

Abstract

The communication outlines the importance of the safeguarding of the Portuguese Language documental heritage. During some conventions, information professionals and researchers jointed to discuss about difficulties and requirements and planned together information strategies, as the Letters of Aveiro and Sao Paulo, the IFLA- Portuguese Caucus, the Lusophone Forum of Archivists and professional training. In 1996, the creation of the Community of Portuguese Language Countries (CPLP), joints eight countries to deepen mutual friendship and cooperation to achieve common objectives. The Informational Policies of the organization structured initiatives for safeguarding of the common cultural heritage, in which documentation as a remarkable position. Till now, most of the digitization and preservation projects are Portuguese and Brazilian. A number of the them are briefly described and they mainly consists in digital programs in different areas of knowledge such as the Health Information Network in Portuguese, ePORTUGUESe, formalized in 2004, and supported by the World Health Organization (WHO), to promote information networks and the creation of the Global Health Library; the Lusophonia Portal, a forum, about the Industrial Property in the Lusophone countries, where you can find, since 2007, a virtual library, the LUSOPAT database and another data of cooperation; the Information Center on Social Protection (CIPS), responds the information needs and experiences of social protection of the CPLP countries, created in 2007, by the Executive Secretary of CPLP, the International Labor Organization (ILO)-Lisbon and STEP Project/ Portugal; the Legis PALOP, born in 2008, containing a platform to provide and share legal information and knowledge among African Countries of Portuguese Official Language (PALOP), developed by the Regional Indicative Program PALOP II, funded by the 9th European Development Fund (EDF) and the Portuguese Institute for Development Support (IPAD). The presentation ends up with a recent initiative, result of a doctoral thesis, containing a methodological guide for the implementation of the Lusophone Digital Library (BDL), a free access portal to digital content of the collections of Lusophone National Libraries. The objective of this contribution is to safeguard the documental heritage, democratize the access to it, increase awareness of its significance and distribute, on large scale, products derivate of it, develop the integration of the Lusophone countries in the knowledge society and promote the Portuguese language in the cyberspace.

Authors

Fernanda Maria Melo Alves studied at the Eduardo Mondlane University in Mozambique, at the Higher Institute of Languages and Administration, and the Universities of Lisbon, Autonomic and Open in Portugal, and at the Carlos III University of Madrid in Spain. She taught Portuguese and French language and culture in Portugal and, in 1998, joined the Information Science Department, at the Carlos III University of Madrid, in Spain, as a researcher, teacher, and manager of international cooperation projects with Lusophone African countries. And she also collaborates with the Cathedra of Portuguese Studies Luís de Camões, at the same university. She has participated in national and international meetings and published in scientific journals. She also works as translator and interpreter, collaborates in NGOs projects, and is president of International Association of the Lusophone Digital Library, located in Portugal.

José Antonio Moreira González has worked in the Universities: UNED (1982-1986), Complutense (1986-1989), Murcia (1989-1991) and Carlos III of Madrid (from 1991). His teaching and investigation attend to the analysis of documentary content, the semantic vocabularies, the documentary theory and the labor market in Information-Documentation. He has participated in two European projects, directed or collaborated in seven of the CICYT, four of the Community of Madrid and five PCI of the AECID. Twenty two doctoral theses were directed by him. He has collaborated like visiting professor in forty universities abroad, seven of them in extended stay.

José Manuel Matias studied History in the University of Lisbon and got an MSc in African Studies at the Lisbon Higher Institute of Social Sciences and Labor (ISTEC). He worked as advisor of the Ministry of Education of Mozambique, from 1975 to 1977. He taught History in secondary education in Lisbon, from 1977 to 1993, later at the Zimbabwe University, from 1993 to 1998, and at the Agostinho Neto University in Angola, from 1998 to 2000. He worked as advisor for cultural and linguistic cooperation with Africa in the Instituto Camões, from 2000 to 2007. He participated in international congresses and symposia about African Studies, and is vice-president of International Association of the Lusophone Digital Library, located in Portugal.

1. Safeguarding in the Digital Age: digitization and preservation planning

All areas of life produce digital documents and the digital documentary heritage, born digital or digitalized traditional documents, is an important heritage for humanity. In order to allow the universal access, the safeguard of the digital documents and digital technology concern everybody, international organizations, governments, experts, professionals and all peoples.

The Memory of the World, launched by UNESCO, seeks to preserve documentary heritage, which reflects the diversity of languages, peoples and cultures and is the mirror of the world and its memory. But this memory is fragile and every day disappears. In order to avoid this situation, the Programme proposes some principles to the preservation of archives, libraries and museums collections all over the world and ensures their wide dissemination.

At the same time, other initiatives improve this policy, like the Information for All Programme (IFAP) in 2006, which looks after a world where everyone has access to information that is relevant to them and where everyone has the opportunity and skills to use this information in creating better societies. The Programme provides a platform for discussions and action on information policies and the safeguarding of recorded knowledge, but many challenges yet remain, such as strengthening the program in all regions, increasing international and intra-regional cooperation to use ICTs for development and materialize its goals.

But resources of information and creative expression are increasingly produced, distributed, accessed and maintained in digital form, creating a new legacy, the digital heritage, that is at risk of being lost and that its preservation, for the benefit of present and future generations, is an urgent issue of worldwide concern. Consequently, UNESCO published in 2003 the *Guidelines for the Preservation of Digital Heritage*,¹ based on the experience a group working in preservation and research programs around the world, supports a variety of preservation strategies and requirements and tools across institutions and settings make the decision on which solution to implement preservation planning.

¹ Guidelines for the preservation of digital heritage (online). UNESCO, Information Society Division, 2003, <http://www.unesco.org/webworld/mdm>.

Besides, the *Charter on the Preservation of Digital Heritage* gives strategies and policies to develop the preservation of the digital heritage, taking into account the level of urgency, local circumstances, available means and future projections. The cooperation of holders of copyright and related rights, and other stakeholders, in setting common standards and compatibilities, and resource sharing, will facilitate this.

A number of initiatives to preserve digital materials have been ongoing in scientific and scholarly research, so that they are maintained and can be re-used. National libraries generally approach the digital environment from the angle of deposit legislation and several libraries have developed strategies for actively selecting and preserving websites, and some national archives have extended policies for electronic record management to include websites of government agencies, public sites and intranet sites, and developed guidelines describing best practices. Other institutions are focusing on collecting materials in a specific subject.

However, as few countries have adopted a national policy regarding digital information and most decision-makers has no sensibility for the problem or are not informed about the risk of disappearance of commonly used means of transmitting and storing digital information.

2. Portuguese Language: Past, Present and Future

Portuguese navigations began by the mid-1500s, and the Portuguese controlled, during centuries, some colonies all over the world.² The Portuguese Empire was motivated by economic, mercantile, political and strategic interests, allied to a certain cultural and scientific curiosity and an attempt at evangelization. The process of colonization encouraged the miscegenation and the colonial education system improved the use of the Portuguese language.³

Following the Second War World, and after the long process of European colonization, most of the European territories gained their independence. But, despite armed struggles, only after the Democratic Portuguese Revolution of 1974, Portuguese colonies finally won independence and the Portuguese language was the major cultural heritage shared between the old metropolis and new nations.⁴

Like Brazil, two centuries before,⁵ the new governments, which came to power in new Lusophone countries, in order to achieve national unification among of numerous ethnic groups, decided that Portuguese is their official or co-official language, cultures and languages, and established cooperation in different areas.⁶

² Boxer, Charles. *O Império Marítimo Português (1415-1825)* (Lisboa: Edições, 1992), 70.

³ Vasconcelos, José Leite de. "História da Língua Portuguesa: origem e vida externa (em linha)," *Revista Lusitana* 25 (1923 a 1925): 5-28.

⁴ Reis, Carlos. La internacionalización de la lengua portuguesa. Contextos, confrontaciones y prioridades (online), 2010. Ciclo de Conferencias 2010. El espacio ibérico de las lenguas. Instituto Cervantes, Madrid, 2010, http://scholar.google.es/scholar?cluster=630868506494526101&hl=es&as_sdt=0.

⁵ República Federativa do Brasil. Portal do Governo Brasileiro Brasil. Português, a língua oficial do Brasil (online), http://www.brasil.gov.br/pais/lingua_portuguesa/portugues.

⁶ Kakunda, Vatomene. "As Línguas Africanas dos Países de Expressão Portuguesa. Passado e presente," in *Forum Internacional de Cultura e Literatura Africanas*, Lisboa: Universidade Católica, 1996.

Portuguese is the official language of the following eight countries, Portugal in Europe, Angola, Cape Verde, Guinea Bissau, Mozambique and Sao Tomé and Príncipe in Africa, and East Timor in Asia, and is spoken by immigrants that live in different regions.⁷

It has been estimated about 240 million people speaking Portuguese in the world.⁸ Portuguese language is the fifth most spoken language on the planet, and the third most spoken language of the western hemisphere, after English and Spanish, and the most spoken language in the southern hemisphere. In the future, Portuguese Language speakers will grow up.⁹

3. The Community of Portuguese Language Countries (CPLP): creation and development

The Community of Portuguese Language Countries (CPLP) is inspired in others similar organizations jointed together around common languages, as English, French, Arab and Spanish, and with similar objectives, and its development is marked by sequential phases.¹⁰

The first one is a large period of the Portuguese colonization that, in the last decades, was fragmented by a new political consciousness recognizing that Portugal must finish with the dictator regime and be a free country, and the African colonies did not belong to Portugal, and the peoples of the colonies had right to freedom and independence.

Those new ideas prepared the Portuguese Democratic Revolution in 1974, the beginning of the second phase, and brought independence to the Portuguese colonies, that choired new ways of organization and development. Unfortunately, military conflicts between rival political groups devastated Angola, Mozambique and Guinea Bissau, and East Timor was invaded by Indonesia till 2002, however the new century bring peace to them, condition for the growth of the peoples. In spite of general problems and difficulties, a proposition of Brazil was accepted, and the Lusophone countries founded in 1989 the International Portuguese Language Institute (IILP).¹¹

The third phase of the CPLP corresponds to a period around the concept of the Lusophonie,¹² a cultural identity shared by eight countries and various communities in the world, united by a common past and language, enriched in its diversity. The concept got official partnership in 1996, with the creation of the Community in Lisbon, and despite a series of delays and various opinions, this moment caused great impact and had great visibility.¹³ The organization is a multilateral forum, privileged to deepen the mutual friendship and cooperation among its members, and has the following objectives:¹⁴

⁷ United Nations, Department of Economic and Social Affairs, Population Division. *World Population Prospects: The 2010 Revision Press Release*, <http://esa.un.org/unpd/wpp/Documentation/publications.htm>.

⁸ Observatório da Língua portuguesa. *Falantes de Português*, <http://observatorio-lp.sapo.pt/pt/dados-estatisticos/as-linguas-mais-faladas/falantes-de-portugues>.

⁹ Internet World Stats. *Top Ten Languages Used in the Web (Number of Internet Users by Language)*, <http://www.internetworldstats.com/stats7.htm>.

¹⁰ Comunidade dos Países de Língua Portuguesa (CPLP), <http://www.cplp.org/>.

¹¹ Instituto Internacional da Língua Portuguesa (IILP), <http://www.iilp.org.cv/>.

¹² Eduardo Lourenço. "Cultura e Lusofonia ou os Três Anéis," in *A Nau de Ícaro seguido de Imagem e Miragem da Lusofonia*. Lisboa, Gradiva, 1999.

¹³ Viggiano, Alan. *Missão em Portugal*. José Aparecido e a Comunidade dos Países de Língua Portuguesa. Brasília: André Quice, ed., 1996.

¹⁴ Comunidade dos Países de Língua Portuguesa. *Estatutos* (online), <http://www.cplp.org/>.

- The conservation of political diplomacy among its Member States, reinforcing its presence in the international scenario.
- The cooperation in all matters, including education, health, science and technology, defense, agriculture, public administration, communication, justice, public security, culture and sports.
- The implementation of projects that promote and disseminate the Portuguese language.

The foundation *Declaration*¹⁵ emphasized the development of an identity based on a common language, the Portuguese, on dialogue with other national languages and on solidarity and cooperation, respecting sovereign equality of states.

The CPLP aims to collaborate in promoting peace, democracy and sustainable development, recognizing multilateral cooperation as the most effective means to achieve these objectives, established above a strong structure in different areas.

4. Safeguarding of the Portuguese Language documentary heritage: approaches to digital preservation and the Lusophone Digital Library

The safeguarding of the Portuguese Language documental heritage, specially, digital documents, had been a preoccupation of the information professionals and researchers. During some conventions, they jointed to discuss about their difficulties and requirements and planned together information strategies, as the Letters of Aveiro and S. Paulo, the IFLA- Portuguese Caucus, the Lusophone Forum of Archivists and professional training.¹⁶

In 1996, the creation of the Community of Portuguese Language Countries (CPLP), joints eight countries to deepen mutual friendship and cooperation to achieve common objectives in different areas. The Information Policies of the organization structured initiatives for safeguarding the common cultural heritage, in which documentation as a remarkable position. Till now, most of governmental and private initiatives of digitization and preservation projects are organized by Portuguese and Brazilian institutions, which have experienced participating in international projects. We present some of them in next paragraphs.

The assignment of the ePORTUGUÊSe¹⁷ was formalized in 2004, during the Ministerial Conference on Health Research in Mexico, and is a model developed by the Latin American and Caribbean Center on Health Sciences information (BIREME) and supported by the World Health Organization (WHO), an organization that promotes information networks and the creation of the Global Health Library.

The project, born in 2004, was created for the following objectives:

- To support the development of human resources for health in the CPLP countries, strengthening collaboration in health information and training.
- To promote the development of the Virtual Health Library.

¹⁵ Comunidade dos Países de Língua Portuguesa (CPLP). *Comunicado Final da Cimeira Constitutiva da Comunidade dos Países de Língua Portuguesa (CPLP)* (online), <http://www.cplp.org/Arquivo.asp?id=5>.

¹⁶ Associação Portuguesa de Bibliotecários, Arquivistas e Documentalistas (BAD), <http://www.apbad.pt/Cooperacao.htm>.

¹⁷ ePORTUGUÊSe, <http://www.bvs.eportuguese.org/php/index.php>.

- To help the dissemination of the Blue Trunk Library in Portuguese, facilitating interaction between health institutions and contributing to the training in health.

Another example is the Lusophonia Portal,¹⁸ a virtual forum where, since 2007, you can find relevant information of the system of Industrial Property of the various Portuguese speaking countries and of the international organizations. To fulfill its mission, this forum aims:

- To contribute to the creation of the necessary conditions for the affirmation of Portuguese as a technology language in the context of the knowledge society.
- To provide a documentation center in Portuguese, which includes the LUSOPAT database, the largest database in the world of patents in Portuguese, a cooperation database to support cooperation initiatives, restructuring and maximizing their results.
- To promote effective partnership about Industrial Property.

The International Labour Organisation (ILO) launched in 2003 the Global Campaign on Social Security and Coverage for All, which conclusions were applied worldwide. In this context, the Project STEP/Portugal, the Ministry of Labor and Social Solidarity of Portugal and the Executive Secretary of CPLP created in 2007 the Information Center on Social Protection (CIPS).¹⁹ The objectives of the CIPS are:

- To provide information in Portuguese to managers with responsibilities related to the social protection, improving knowledge and information in this area.
- To guide policy decisions, supporting decision makers through research and experience and provide a framework of different approaches and strategies used in the Portuguese-speaking world.
- To create opportunities for institutions and civil society, enhancing the exchange of information and knowledge about the extent of social protection for the benefit of the people.

In 2010, in Fortaleza, the Ministers of Labor and Social Affairs of the CPLP recognized the importance of the CIPS and adopted the resolution to ensure the continued development of the center.

In the legal context, is worthy of reference another case, the Legis-PALOP,²⁰ inserted under the Project Support Development to Judicial Systems of the PALOP (Angola, Cape Verde, Guinea Bissau, Mozambique S. Tomé and Príncipe), inserted in the PALOP Regional Indicative Programme II, financed by the 9th European Development Fund (EDF) and the Portuguese Institute for Development Support (IPAD).

The project, initiated in 2008, is an ambitious project of providing a platform of knowledge and of legal information among the African Countries of Portuguese Official Language (PALOP) and for those who wants to know these subjects, providing access to legislation, jurisprudence (higher courts) doctrine and documents, published after the independence and a common thesaurus. Besides, in partnership with

¹⁸ Portal da Lusofonia, <http://www.portal-lusofonia.org/>.

¹⁹ Centro de Informação em Proteção Social (CIPS), http://www.cipsocial.org/index.php?option=com_content&task=view&id=57&Itemid=105.

²⁰ Legis-PALOP, <http://www.legis-palop.org/bd>.

the Overseas Historical Archive of Institute of Tropical Research (IICT), were collected all legislation published since 1926 to the African independences.

The Legis-PALOP became a recognized instrument of Lusophone citizenship and civic and economic development, actively contributing to the consolidation a fairer society and a more solid international position of the PALOP.

In this context, the presentation ends up with another recent initiative, consequence of a doctoral thesis,²¹ defended in 2007. The research emphasizes the privileged place of the national library, as an encyclopedic institution, for the conservation of the heritage literature produced in a country, the responsibility to elaborate the national bibliography and the obligation to provide the national heritage to the largest possible number of users.²²

The economic growth provided the increased of literacy of the population and the demand of the information. National libraries were obliged to enrich their collections, introduce new areas of knowledge, practice new functions and question their new mission and priorities in a context of the new digital world.²³ Besides, the National Library is one of the key elements of the national information policy, as well as the national archives and the archaeological museum, serving the preservation and dissemination of cultural heritage of each country.²⁴

In this perspective, the study of the Lusophone National Libraries allowed to design a theoretical, conceptual and methodological proposal, called *Methodological Guide for the Implementation of Lusophone Digital Library*, which the potential users exceed 240 million Lusophone people.

The mission of this initiative is the establishment of a cooperation project, oriented to the collection, access, preservation and distribution of a network of digital objects belonging to cultural and documental heritage of the Lusophone National Libraries, as well as the creation of their infrastructure and services.

The proposal includes the creation of a free access portal to digital content, contributing for the integration of the Lusophone countries in the knowledge society, the safeguarding of the documental heritage, the democratization of the access to it and promotes the Portuguese language in the cyberspace.

The International Association of the Lusophone Digital Library, in Portugal, the Lusophone National Libraries and the Information Science Department of the Carlos III University of Madrid, in Spain, recently developed a project for the safeguarding of the Portuguese Language documentary heritage, named the Lusophone Digital Library, which is waiting for financial support of the UNESCO.

²¹Melo Alves, Fernanda Maria. Articulación y complementariedad de las políticas de la lengua portuguesa, de cooperación y de información en los países lusófonos: guía metodológica para la implantación de la Biblioteca Digital Lusófona (BDI) (online), <http://e-archivo.uc3m.es/handle/10016/2/browse?type=author&order=ASC&rpp=20&value=Alves%2C+Fernanda+Maria+Melo>.

²² Carrión Gútiérrez, Manuel. *Manual de Bibliotecas*, 2ª (Madrid: Fundación Germán Sánchez Ruipérez, 1993), 13.

²³ Line, Maurice. "The role of national library: A reassessment," *Libri* 30, no. 1 (1980): 1-16.

²⁴ Guinchat, Claire, Menou, Michel. "La gestion et les politiques d'information au niveau national et international," in *Introduction Général aux Ciencias et Técnicas de la Información et de la Documentación*, ed. C. Guinchat and M. Menou (Paris: UNESCO, 1981), 369-370.

Le réseau francophone numérique¹

Benoit Ferland et Tristan Müller

Bibliothèque et Archives nationales du Québec (BAnQ)

Résumé

Le Réseau francophone numérique (RFN) offre un accès centralisé au patrimoine documentaire francophone numérisé. Le Réseau rassemble aujourd'hui 24 institutions. Le portail du RFN diffuse gratuitement des centaines de milliers de documents en langue française. Plusieurs activités de formation et de transfert de savoir-faire ont été réalisées auprès d'un nombre croissant d'institutions documentaires de la Francophonie. Pour les institutions patrimoniales francophones membres, le réseau constitue également un forum d'échanges autour des enjeux de conservation du patrimoine documentaire à l'ère numérique.

Auteurs

Benoit Ferland est le directeur général de la conservation à Bibliothèque et archives nationales du Québec. Il est détenteur d'une maîtrise de l'École de bibliothéconomie et des sciences de l'information de l'Université de Montréal (EBSI), d'une maîtrise en administration publique (ENAP), d'un certificat en archivistique (UQAM) et d'un diplôme de premier cycle en histoire (Université de Montréal). En plus d'être chargé de cours à l'EBSI, M. Ferland agit régulièrement en tant que consultant et formateur. Il œuvre depuis près de 30 ans dans le milieu des bibliothèques. Il est l'auteur d'un ouvrage de bibliothéconomie : *L'élaboration des politiques en milieux documentaires*.

Tristan Müller a obtenu un diplôme de maîtrise en bibliothéconomie et en sciences de l'information de l'Université de Montréal en 1999. Il a travaillé au Jardin botanique de Montréal, puis pour un projet de la Banque Mondiale au Burkina Faso. Depuis 2004, il travaille à Bibliothèque et Archives nationales du Québec (BAnQ); d'abord comme bibliothécaire, puis comme chef du service du prêt. Il est directeur de la numérisation depuis 2012.

1. Introduction

Depuis le début des années 2000, plusieurs projets internationaux de bibliothèques numériques ont vu le jour : Bibliothèque numérique mondiale (World Digital Library), Europeana, Google Books, etc. Une rapide analyse du contenu de ces bibliothèques montre que la très grande majorité des documents diffusés sont en langue anglaise. La raison en est fort simple : les principaux fournisseurs de documents numériques proviennent de pays anglo-saxons. Pourtant, le patrimoine documentaire mondial n'est pas limité à l'anglais. Le français, par exemple, est une langue parlée par plus de 220 millions de personnes à travers le monde² et l'une des deux seules langues parlées (avec l'anglais) sur les cinq continents. Chaque année, plus de 70 000 livres³ sont publiés dans cette langue. Or, à l'aube du 21^e siècle, le constat est frappant : le patrimoine documentaire de la Francophonie est pratiquement absent de ce nouveau paysage

¹ Réseau francophone numérique (RFN), Bibliothèque et Archives nationales du Québec 475, boul. De Maisonneuve Est, Montréal (Québec) H2L 5C4. info@rfnum.org, <http://www.rfnum.org>

² <http://www.francophonie.org/Qui-sommes-nous.html?test=mobile>.

³ http://www.dgmic.culture.gouv.fr/IMG/pdf/Chiffres-cles_2010-2011.pdf.

numérique. L'omniprésence de la langue anglaise sur le Web fait craindre, à plusieurs, que la présente situation conduise à une lente, mais inéluctable uniformisation culturelle.

2. Historique

À la lumière de ce constat inquiétant, l'Organisation internationale de la Francophonie insiste dès 2004 pour que le développement durable développement soit « attentif à la diversité culturelle et linguistique. »⁴ Dans cette mouvance, la Bibliothèque nationale de France a lancé, en 2006, un appel aux bibliothèques nationales francophones pour une diffusion concertée du patrimoine documentaire francophone sur le Web. Six bibliothèques nationales de la Francophonie (Belgique, Canada, France, Luxembourg, Québec, Suisse) ont répondu à l'appel. Toutes ont reconnu le rôle crucial des programmes de numérisation pour le rayonnement des cultures francophones et de la langue française. Pour relever ce défi, ce groupe se constitue alors en un Réseau francophone des bibliothèques nationales numériques (RFBNN). Peu de temps après l'Égypte, qui est membre de la francophonie, s'est joint au groupe.

Compte tenu de son importance aux yeux des dirigeants francophones, le nouveau réseau reçoit d'emblée l'appui⁵ de l'Organisation internationale de la Francophonie (OIF⁶). Cette organisation est composée de 75 États et gouvernements (56 membres et 19 observateurs) — soient plus du tiers des États membres des Nations unies. En 2007, dans un discours aux bibliothèques nationales et patrimoniales des pays ayant le français en partage, Abdou Diouf, Secrétaire général de la Francophonie, cerne assez bien l'enjeu stratégique de la présence du fait français sur Internet :

Ce qui se joue, en ce moment même, c'est la présence de la langue française, l'existence des cultures en langue française dans l'espace numérique. Demain, ce qui ne sera pas numérisé et rendu accessible en ligne, risque d'être tout simplement occulté, pour ne pas dire oublié. Or, notre communauté a de grandes richesses à partager et à faire partager.⁷

Afin d'offrir un accès centralisé au patrimoine documentaire francophone numérisé, les membres du Réseau francophone des bibliothèques nationales numériques décident de mettre en place un portail Web. La conception et la réalisation technique de ce portail ont été confiées à Bibliothèque et Archives nationales du Québec (BAnQ). En 2008, le portail Web du RFBNN est officiellement lancé à l'occasion du XIIe Sommet des chefs d'État et de gouvernement de la Francophonie à Québec.

En 2010, le Réseau s'ouvre à toutes les institutions chargées de préserver et de diffuser le patrimoine francophone. C'est ainsi que le RFBNN change de dénomination et devient le RFN : le Réseau francophone numérique. En 2012, en à peine six ans d'existence, le RFN est passé de 6 à 24 institutions. On peut parler d'une croissance importante et soutenue.

Il faut par ailleurs souligner que l'Organisation internationale de la Francophonie participe activement au RFN, notamment en jouant un rôle d'observateur lors des diverses rencontres des membres du RFN, mais également en offrant d'autres formes d'appuis (financier, logistique, etc.).

⁴ <http://www.tlfq.ulaval.ca/axl/francophonie/ouadagoudou2004.htm> — Déclaration de Ouadagoudou.

⁵ XIe sommet de la Francophonie — Déclaration de Bucarest.

⁶ <http://www.francophonie.org/>.

⁷ Discours du 13 septembre 2007, Bruxelles.

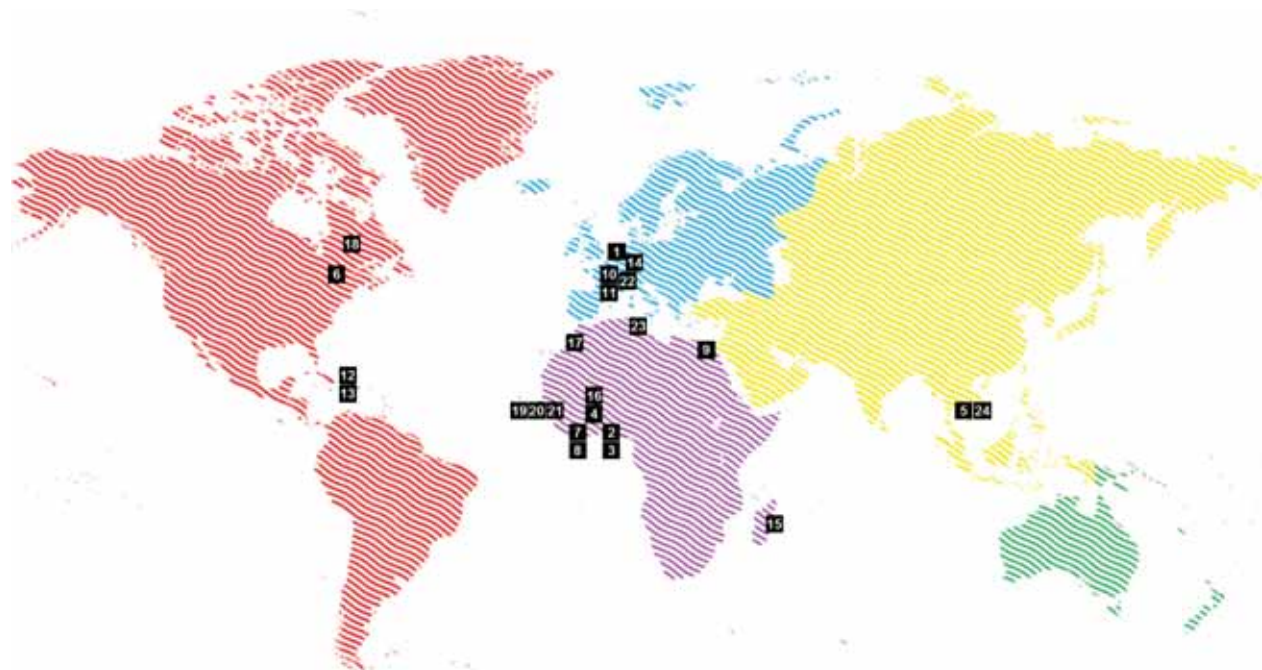


Figure 1. Carte des 24 institutions membres du RFN.

Tableau 1. Liste des institutions membres du RFN.

1. Bibliothèque royale de Belgique
2. Bibliothèque nationale du Bénin
3. Archives nationales du Bénin
4. Bibliothèque nationale du Burkina Faso
5. Bibliothèque nationale du Cambodge
6. Bibliothèque et Archives Canada
7. Bibliothèque nationale de Côte d'Ivoire
8. Archives nationales de Côte d'Ivoire
9. Bibliotheca Alexandrina (Égypte)
10. Bibliothèque nationale de France
11. Bibliothèque francophone multimédia de Limoges
12. Bibliothèque nationale d'Haïti
13. Bibliothèque haïtienne des Pères du Saint-Esprit
14. Bibliothèque nationale de Luxembourg
15. Bibliothèque universitaire d'Antananarivo (Madagascar)
16. Bibliothèque nationale du Mali
17. Bibliothèque nationale du Royaume du Maroc
18. Bibliothèque et Archives nationales du Québec
19. Bibliothèque des Archives nationales du Sénégal
20. Bibliothèque centrale de l'Université Cheikh Anta Diop (Sénégal)
21. Institut fondamental d'Afrique noire (Sénégal)
22. Bibliothèque nationale suisse
23. Bibliothèque nationale de Tunisie
24. Bibliothèque nationale du Vietnam

3. Gouvernance

Chaque année, une assemblée générale de tous les membres du RFN est organisée. L'assemblée adopte les clauses de la charte du Réseau, propose un président d'assemblée, nomme un comité de pilotage, approuve l'adhésion de nouveaux membres et se prononce sur les développements et les orientations du Réseau.

Le comité de pilotage, composé de 7 membres et d'observateurs de l'OIF, se réunit deux fois par année et joue le rôle de comité exécutif. Il définit, entre autres, les grandes orientations et les axes de développement du portail du RFN et assure l'engagement et la participation continue et soutenue des institutions membres.

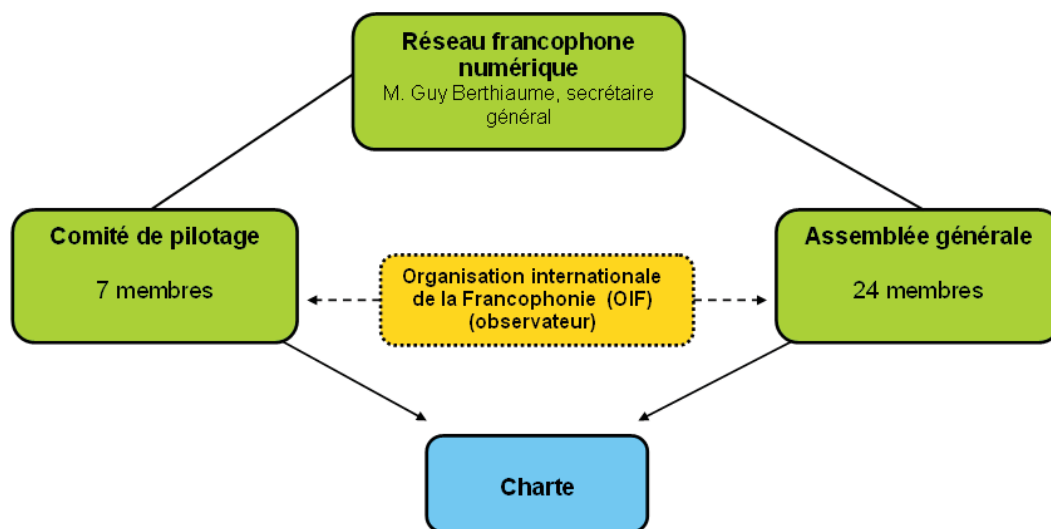


Figure 2. Structure de gouvernance du RFN.

4. Mission

Le RFN s'est donné une triple mission:

1. Grâce à la numérisation, **préserver un patrimoine précieux** souvent menacé de disparition et le diffuser gratuitement auprès d'un large public.
2. Assurer le **transfert de savoir-faire** auprès d'un nombre croissant d'institutions documentaires de la Francophonie.
3. Offrir aux institutions patrimoniales de l'espace francophone un **forum d'échanges** autour des enjeux de l'ère numérique.

Par ce très louable énoncé mission, le RFN s'engage en faveur de la diversité culturelle, du rayonnement de la culture francophone et de la solidarité internationale. Pour les pays du Sud, il s'agit d'une incitation à préserver leurs documents rares et précieux grâce à la numérisation. Enfin, le RFN est considéré comme un moyen unique de mise en valeur et de diffusion du patrimoine documentaire francophone dans le monde entier.

5. Activités

5.1. Portail RFN

On peut affirmer que la préservation des documents patrimoniaux profite grandement des avancées technologiques des outils de numérisation. Les documents patrimoniaux numérisés permettent de protéger numériquement et souvent à distance les documents originaux. La diffusion de ce patrimoine numérisé passe principalement par le web.⁸ Il s'agit d'une vitrine exceptionnelle vers les collections francophones numérisées. L'accès à la richesse des diverses collections nationales, déjà diffusées sur les sites Internet des institutions membres du Réseau, se trouve multiplié par le regroupement sur un même portail. Il faut en effet considérer que les divers pays membres, même s'ils partagent un attachement à la langue française, présentent des éléments culturels propres qui sont autant de facettes culturelles spécifiques qui enrichissent la diversité culturelle de la Francophonie.

Le portail du RFN donne accès à des collections de journaux, revues, livres, cartes et plans, archives et enregistrements vidéos numérisés par 15 pays. Les principes qui prévalent au développement du portail du RFN sont:

- Utiliser les collections numérisées existantes et établir des liens avec les sites web des institutions membres;
- Alimenter progressivement le portail de nouvelles collections de documents;
- Assurer un accès gratuit aux documents.

En 2012, le portail donne accès à :

Tableau 2. Documents accessibles par le portail du RFN en 2012

Journaux	Plus de 640 000 fascicules 75 titres Essentiellement du 19e et 20e siècle
Revue	Près de 3 700 fascicules 39 titres Principalement du 20e siècle
Livres	Environ 40 titres
Cartes et plans	Collection thématique America, 1500-1800 (BAnQ) Le Canada à l'échelle : les cartes de notre histoire (BAC)
Archives	7 ensembles d'archives Ex. : Mémoires du Canal de Suez (Bibliotheca Alexandrina)
Documents audiovisuels	Le Cercle littéraire et conférences en ligne de la BnF Arthur Lamothe et les Innus : culture amérindienne (BAnQ) Au fil des mots : treize poètes québécois se racontent (BAnQ)

Il est pertinent de préciser que le portail du RFN ne contient pas de documents numériques, mais plutôt des listes ou index de métadonnées des collections numériques.

⁸ <http://www.rfnum.org>.

Concrètement, l'ajout d'un nouveau document au portail consiste à extraire les métadonnées disponibles sur le site Web de l'institution et de les transformer conformément au Dublin Core qualifié. Cette opération se fait actuellement par la récupération directe via FTP de fichiers de métadonnées (ex.: XML) ou par la génération manuelle de fichiers de métadonnées. Celles-ci sont ensuite transformées en fichier XML puis chargées dans Solr.⁹ Le portail constitue en définitive une sorte d'entrepôt de métadonnées des collections numérisées des institutions membres du RFN. Ce procédé permet de pointer directement sur les documents numérisés, donnant ainsi un autre point d'accès aux internautes du monde entier intéressés par des documents de langue française. À son tour, le portail du RFN peut être moissonné par d'autres bibliothèques numériques puisqu'il utilise le protocole OAI-PMH.¹⁰

Le moteur de recherche Solr, intégré dans le logiciel de gestion de contenu web dotCMS,¹¹ permet de faire une recherche dans les métadonnées. L'interface actuelle offre des possibilités de recherches par type de document, pays, titres et dates. La conception, l'hébergement et la gestion du portail RFN sont effectués par BAnQ, en concertation avec la BnF, la Bibliotheca Alexandrina et l'OIF.

Enfin, le portail comprend aussi un espace professionnel dédié à la numérisation. Dans cette section, on retrouve des documents de formation réalisés pour les stages et d'autres documents complémentaires sur les méthodes et normes de numérisation.

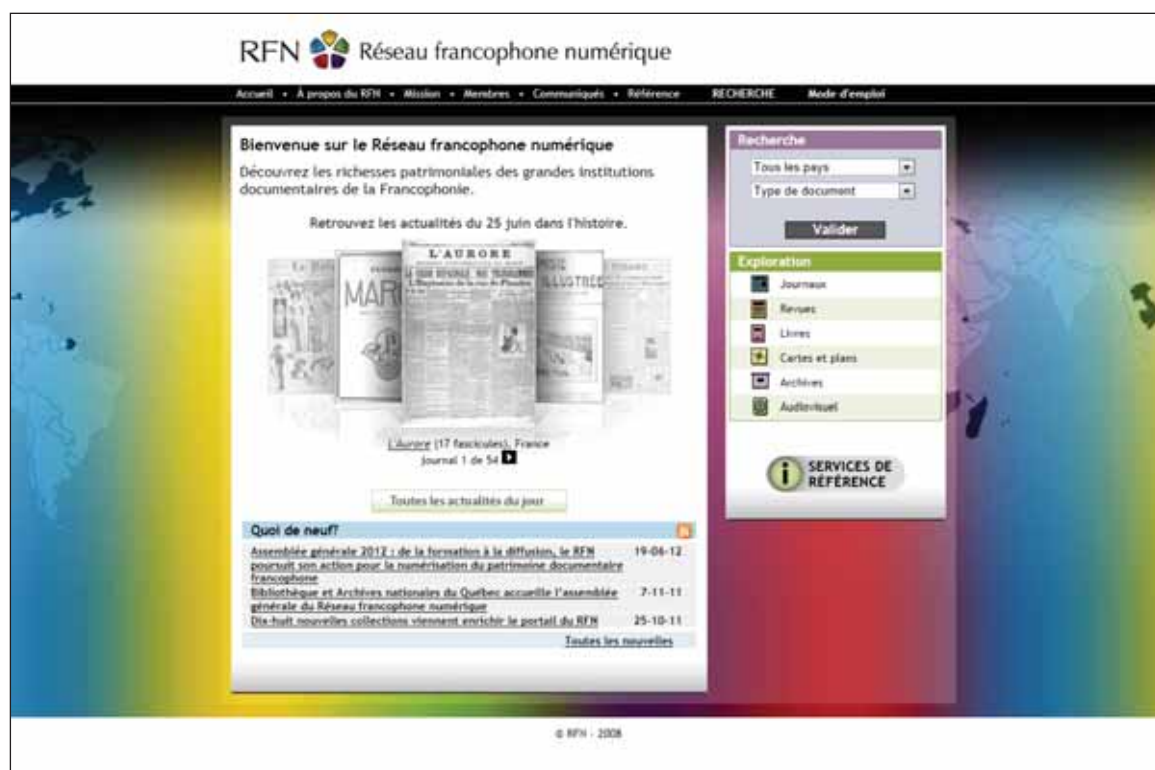


Figure 3. Interface du portail RFN.

⁹ <http://lucene.apache.org/solr/features.html>.

¹⁰ <http://www.openarchives.org/pmh/>.

¹¹ <http://dotcms.com/>.

La mise en œuvre du portail comporte plusieurs défis. Il faut comprendre que la qualité des documents fournis par les institutions est inégale. Les institutions ne bénéficiant pas toutes des appareils derniers cris. Il appert en outre que les métadonnées peuvent être parfois incomplètes ou carrément inexistantes. On peut aisément comprendre qu'il s'agisse d'une problématique importante, considérant que le portail utilise les métadonnées comme index référant vers les documents. Une autre difficulté rencontrée est celle de la diversité des standards et des protocoles. Finalement, il faut mentionner que l'accès aux contenus est tributaire de la disponibilité des systèmes documentaires locaux. Toutefois, des solutions existent déjà pour répondre à ces défis. Il reste évidemment à toutes les mettre en œuvre. C'est sans doute par le partage des forces et des expertises des membres du Réseau que ces défis sauront être relevés.

5.2. Stages et mission de formation

Le partage du savoir-faire est une partie essentielle de la mission du RFN. Ce partage des connaissances et des ressources, notamment en ce qui a trait à la numérisation, passe par la coopération entre les pays du Nord et du Sud. Depuis 2008, ce partage a pris la forme de stages de formation tenus notamment en France (avril 2008), en Haïti (juin 2009), au Sénégal (janvier 2011) et au Maroc (mai 2012). Il faut souligner que la participation à ces stages n'étaient pas uniquement réservés au personnel des institutions hôtes, mais aussi aux autres institutions. Par exemple, lors du stage au Maroc, plus de 19 personnes ont été formées. Ceux-ci provenaient du Maroc, de Tunisie, du Sénégal, du Mali, du Burkina Faso, du Togo, du Bénin et de Côte d'Ivoire. Le programme de formation comportait les éléments suivants:

1. Principes de constitution d'une collection numérique et modalités de développement d'un projet.
2. Éléments techniques de la numérisation des documents imprimés, iconographiques et audiovisuels : normes, matériel, logiciels, traitement des fichiers numériques, reconnaissance optique de caractères, etc.
3. Archivage et diffusion des documents numériques : supports et métadonnées.
4. Chaîne et techniques de numérisation : tri et sélection des fonds à numériser, gestion des droits d'auteur, préparation et numérisation des documents (papier et audiovisuels), reconnaissance optique de caractères, contrôle de la qualité des documents numérisés, sauvegardes des documents numérisés sur les serveurs et/ou DVD, création des métadonnées, mise en consultation des documents numérisés.

À l'occasion de ces stages, l'institution hôte se voit offrir des équipements de numérisation (ordinateurs, numériseurs, logiciels et imprimantes) complémentaire à ceux déjà en place. Cette façon de faire permet aux institutions bénéficiaires de disposer d'une plateforme de numérisation très fonctionnelle. Par exemple, lors du stage au Sénégal, un appareil photo Canon D9 ainsi que deux numériseurs à plat ont été acquis. Ceux-ci ont été utilisés pour numériser des documents souvent difficiles à manipuler, notamment les manuscrits du premier président du Sénégal indépendant Léopold Sédar Senghor. Tous ces stages ont été donnés soit par la Bibliothèque nationale de France, Bibliothèque et Archives nationales du Québec ou conjointement et toujours avec la complicité de l'institution hôte. Réunissant des institutions de différents pays, ces stages constituent l'occasion par excellence de discuter sur les pratiques de numérisation. Il s'agit non seulement d'une opportunité de formation, mais aussi d'échanger librement sur divers aspects de la numérisation du patrimoine documentaire de diverses nations dans une langue commune de partage.

En 2009, une mission de numérisation auprès de 3 institutions patrimoniales khmères (Bibliothèque nationale, Musée national, Centre d'études khmères) a eu lieu dans la tradition de partage du savoir-faire du Réseau. La mission khmère consistait à sélectionner les fonds de périodiques en langue française pertinents à numériser pour le RFN, à examiner les conditions de leur numérisation et à former le personnel permettant la mise en place d'une chaîne efficace de numérisation. La mission a été réalisée par la Bibliothèque nationale de France.

L'impact de ces stages et mission a été immédiat et concret. On a pu constater une amélioration certaine des processus de numérisation en place, et ce, tant en efficacité qu'en qualité. Le transfert de connaissances techniques dans le domaine de la numérisation s'est opéré à un niveau très encourageant. Par la suite, la méthode de formation des formateurs permet aux gens formés lors des stages, d'eux-mêmes formés par la suite leurs compatriotes.

Dans la mise en œuvre du transfert de savoir-faire, plusieurs défis ont été rencontrés : les logiciels et le matériel utilisés pour le calibrage du matériel de numérisation (écrans et numériseurs) ont un coût très élevé pour les institutions du Sud; plusieurs institutions n'ont pas les compétences informatiques ni l'infrastructure technologique pour diffuser leurs documents numérisés dans des bibliothèques numériques (Greenstone,¹² DSpace,¹³ etc.); et certaines institutions qui ont bénéficié d'une formation ne disposent pas toutes des équipements de numérisation pour mettre en pratique les compétences acquises.



Figure 4. Session de formation au Maroc.

6. LE RFN DE DEMAIN

Le Réseau francophone numérique compte actuellement 24 membres et représente 21 pays. Or, l'Organisation internationale de la Francophonie comprend 75 États et gouvernements et le nombre de membres potentiel dépasse la centaine. L'orientation du RFN sera donc de continuer son expansion dans d'autres régions de l'espace francophone, notamment l'Europe de l'Est (ex. : Roumanie) et l'Amérique du Sud (ex. : Guinée française).

La croissance du réseau permettra de multiplier les échanges et les possibilités de parrainage. Le portail du réseau bénéficiera aussi de l'arrivée de nouveaux membres qui ajouteront leurs collections patrimoniales à celles déjà présentes.

¹² <http://www.greenstone.org/>.

¹³ <http://www.dspace.org/>.

L'interface et les fonctionnalités du portail du RFN seront revues en profondeur. Une nouvelle interface plus épurée mettra davantage en valeur les collections des institutions membres. Cette valorisation des collections numériques passera notamment par exergue une mise en valeur des trésors nationaux. Un document reconnu comme un « trésor national » devra présenter une importance historique ou culturelle significative ou encore posséder un caractère unique ou original. En plus de cette mise en valeur, les membres du RFN s'entendent pour enrichir le portail de collections audiovisuelles et de documents patrimoniaux issus de la tradition orale. On prévoit aussi l'ajout de nouvelles fonctionnalités : le développement d'une option de récupération des métadonnées par moissonnage, si l'institution membre possède ce service (ex. : OAI-PMH); des fonctions de recherche simple et avancée; la possibilité de rechercher en texte intégral dans les collections du RFN; un affinement des résultats de recherche par

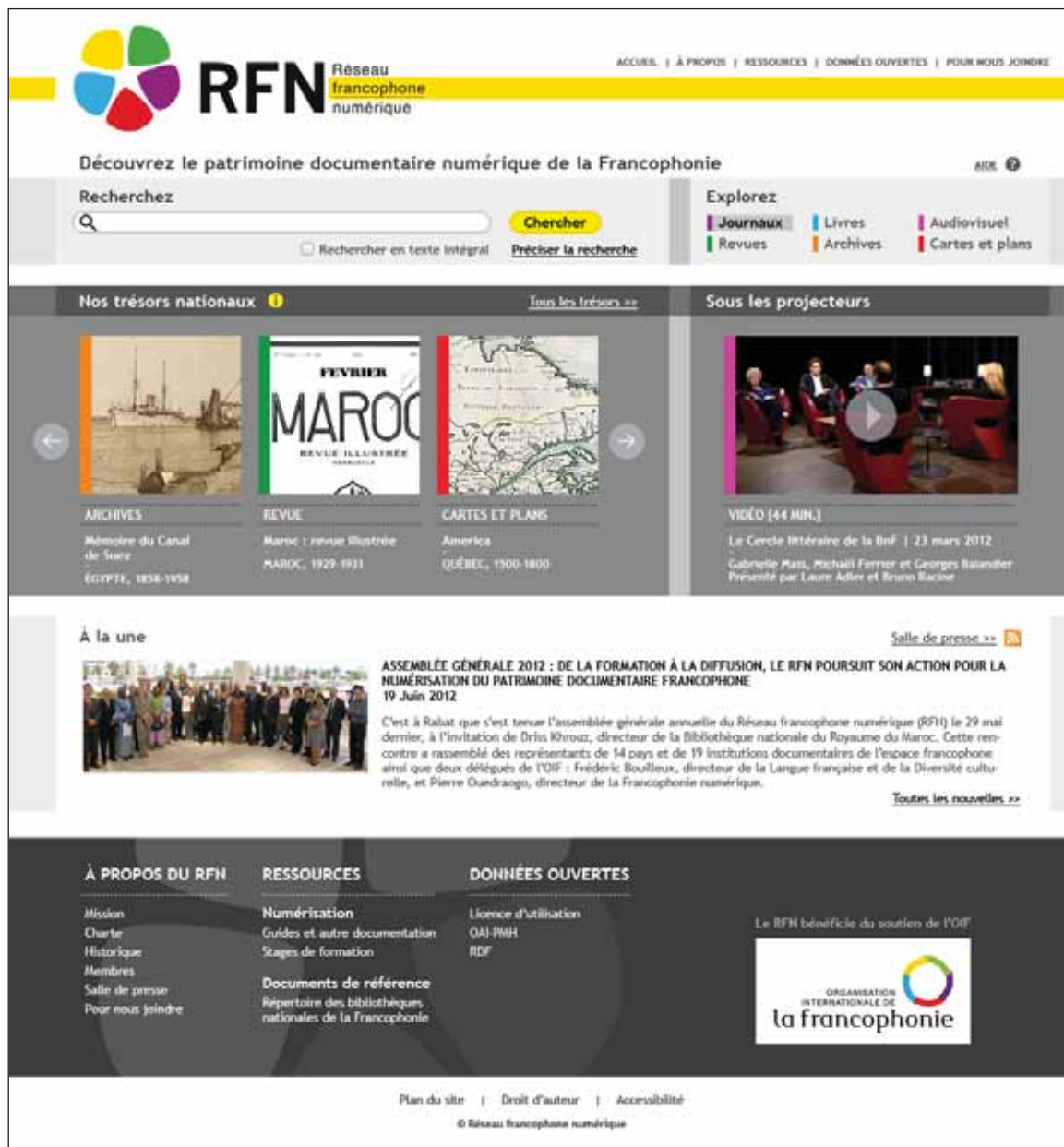


Figure 5. Nouvelle interface du portail RFN.

facettes; des notices plus détaillées; et l'ouverture des métadonnées en mode *Open Access* afin qu'elles puissent être réutilisées à différentes fins.

Afin de combler certaines lacunes informatiques et technologiques de certains membres du RFN, la BnF proposera un nouveau service appelé « Gallica Marque Blanche ». Ce service offrira des possibilités de stockage, de diffusion et d'archivages des documents numérisés sur les serveurs de la BnF. Ce service est personnalisable, c'est-à-dire que chaque bibliothèque numérique aura sa propre identité visuelle adaptée aux couleurs de son institution. Le portail du RFN n'aura qu'à moissonner directement les métadonnées de ces serveurs pour effectuer une nouvelle valorisation de ces collections.

En ce qui concerne le partage du savoir-faire, de nouvelles avenues de transmission sont envisagées. On peut penser notamment à la création de stages régionaux de formation spécialisés en lien avec un projet particulier de numérisation. Un accompagnement sera également assuré pour la mise en œuvre de projets de numérisation, de la sélection du projet jusqu'à la mise en ligne des contenus et leur intégration dans le portail du RFN. Cette méthode permettra d'agir beaucoup plus directement avec le personnel des institutions membres. Ce soutien sera essentiellement assuré à distance après une intervention ciblée auprès d'une institution.

7. Conclusion

Après seulement quelques années d'existence, le Réseau francophone numérique est devenu un formidable moyen de sensibiliser les responsables culturels des États de la Francophonie à l'importance et parfois même l'urgence de sauvegarder leur patrimoine documentaire et de le rendre accessible à tous grâce à la numérisation. L'adhésion à ce Réseau, qui ne cesse de grandir, constitue la preuve concrète qu'une coopération efficace est possible entre institutions, bibliothèques et centres d'archives dont la langue de partage est le français. La diversité culturelle de l'espace numérique se trouve certainement renforcée par la présence du RFN sur le Web. Toutefois, il faut se rappeler que ce type de projet est tributaire de l'engagement actif de toutes les grandes institutions documentaires de la Francophonie et même au-delà. Par exemple, une coopération avec l'UNESCO pour l'accès aux contenus du RFN à partir de la Bibliothèque numérique mondiale (<http://www.wdl.org>) serait probablement bénéfique pour une meilleure diffusion de ce patrimoine mondial.

The World Digital Library

John Van Oudenaren¹

Director, World Digital Library, The Library of Congress, USA, www.wdl.org

Abstract

The World Digital Library is an international collaborative project that fosters access to documentary heritage through digitization, capacity building and technical assistance, and the maintenance and continuous development of www.wdl.org, a website that provides free, multilingual access to cultural heritage. Topics covered in this paper/presentation include: (1) project objectives; (2) content selection and potential legal, ethical, and political issues associated with selection; (3) capacity building challenges in developing countries; (4) partnership with the UNESCO Memory of the World program; (5) understanding and targeting audiences and user groups; (6) the challenges of multilingualism; and (7) the WDL and preservation (physical and digital). Particular attention is devoted to cultural and professional challenges, such as those identified in the conference scope paper, and how the WDL is coping with these challenges, and to the need for coordination among global, regional, national, and institutional projects to meet user demands and to meet the broader global challenge of preserving the “memory of the world in the digital age.”

Author

John Van Oudenaren directs the World Digital Library (www.wdl.org), a collaborative initiative of the Library of Congress, UNESCO, and libraries and other cultural institutions from around the world. Previously he was chief of the European Division at the Library of Congress and director of the Library's *Global Gateway* international digital library collaborations. Prior to joining the Library in 1996, he was a senior researcher at RAND in Santa Monica, California. He received his Ph.D. in Political Science from the Massachusetts Institute of Technology and his A.B. in Germanic Languages and Literature from Princeton University.

1. Introduction

It is now more than seven years since Librarian of Congress James H. Billington first proposed the establishment, in a June 2005 speech at Georgetown University to the U.S. National Commission for UNESCO, of a World Digital Library. At the time, the United States was in the process of rejoining UNESCO after a nearly twenty-year absence. The U.S. national commission was soliciting projects that the United States might bring to the organization. Billington called for a cooperative project, to be undertaken by the Library of Congress and partner libraries from around the world in cooperation with UNESCO, to digitize and make freely available over the Internet primary source documents that tell the stories and highlight the achievements of all countries. Such a project, he argued, “would hold out the promise of bringing people closer together precisely by celebrating the depth and uniqueness of different cultures in a single global undertaking.”²

¹ The opinions expressed in this paper are personal and do not express the views of the Library of Congress or the United States government.

² James H. Billington, The Librarian of Congress, Remarks to the Plenary Session, The U.S. National Commission for UNESCO, Georgetown University, June 6, 2005. Text at: <http://www.loc.gov/about/librarianoffice/speeches/060605.html>

In December 2005, the Library of Congress announced that it had received initial financial support from Google, Inc. to begin planning the WDL. The announcement of funding led to a flurry of interest in the WDL. UNESCO welcomed the initiative, but understandably had questions about how it fit with the organization's program and priorities. The Governing Board of the International Federation of Library Associations and Organizations (IFLA) was briefed on the initiative and expressed interest. A number of heads of national libraries offered to participate. But there was also skepticism. European libraries were embarking on their project to create a European digital library, and some wondered how a WDL would mesh with this effort. In developing countries—and not just developing countries—there were the usual suspicions about American cultural imperialism. Many librarians and academics in the United States and Europe saw the WDL's initial association with Google, even if limited to a no-strings-attached financial contribution, as problematic. At the very least, some in the library community saw in the declared intention to create a *world* digital library a certain over-ambitiousness, a grandiosity that they found off putting.

Within the Library of Congress, a team was established to begin the internal deliberations and external consultations aimed at translating Billington's vision into reality. Exactly what that reality would be was as yet wide open. General objectives had been set, but decisions needed to be made about technical architecture, content selection, target audience, modes of participation, and much else.

This paper will review the main decisions that the Library of Congress team and its leading external collaborators made in planning the WDL and the implications those decisions had for the subsequent development of the project. Its purpose is not to recount history for its own sake, but to draw lessons with the goal of applying these lessons to the overall session topic: "Beyond Access: Digitization to Preserve Culture."

2. Decisions and Choices

The WDL planning team was aware of the skepticism in some quarters about the idea of a world digital library. Indeed, some at the Library of Congress who had participated in various national and international digital library projects had reasons of their own to wonder whether such an ambitious undertaking could succeed. So they proceeded cautiously and incrementally. They decided that the first question to be answered was not, "how do we create a world digital library?" but rather "What should a world digital library look like in order to be worthy of the name?" What capabilities, features, and content should be offered to create a digital library that people will actually find useful? What added value must it offer to attract users who already have access to Google? And how should it differ from or complement the array of national and international digital library projects to which people already have access?

It was also important to plan the WDL with reference to evolving user expectations and emerging technologies. By 2006, when work on the project got underway, the World Wide Web was thirteen years old. At the Library of Congress, projects such as *American Memory* had been underway for a decade. A series of bilateral international digital library collaborations, beginning with the *Meeting of Frontiers* project with Russia, went back to 1999.³ A growing body of evidence, both anecdotal and statistical, existed about what Internet users liked about the digital library projects that cultural institutions such as the Library of Congress were undertaking—but also about what they disliked and found frustrating.

³ This project remains online, and can be seen at <http://frontiers.loc.gov>.

The applications and devices by which people accessed online content also were changing. The introduction of smart phones and the explosion in use of eReaders were just over the horizon (Steve Jobs introduced the iPhone in January 2007, Amazon the Kindle in November of that year), but Google had already revolutionized the world of search. *American Memory* had been designed to enable a student looking, for example, for photographs of the American Civil War to go to the Library of Congress website, to proceed from there to *American Memory*, to look within *American Memory* for the photograph collections, to choose from among the photograph collections that of Mathew Brady, and then to search or scroll through a browse list to find the desired picture. This hierarchical approach, which had grown out of the early attempt to mirror in cyberspace the internal structures of the Library and its custodial units and collections, was now obsolete. Users wanted to type “Civil War soldier pictures” into a search box and be taken directly to the photos they were seeking. This and other changes in user behavior needed to be factored into the planning.

And, as if Google had not already caused enough havoc in the library world, what was at the time called “Google Print” was announced in November 2004 at the Frankfurt Book Fair. A month later, the first set of agreements between Google and leading research libraries—Harvard, the University of Michigan, the New York Public Library, Oxford, and Stanford—for what would become Google Books was announced. This soon led to talk about “every book in the world” being digitized, which at least for some observers created further confusion with a projected “world digital library.”

Against this background, the planners came to focus on three main sets of requirements for the proposed WDL: (1) multilingualism; (2) universality; and (3) a high level of functionality and added value aimed at actual users.

Multilingualism. A true WDL had to be multilingual, with regard both to content and to access. The Library of Congress alone collects in more than 400 languages, and there was no reason why each of these 400 languages (and possibly many others), including endangered languages of particular interest to UNESCO, should not be represented on the WDL.

Access also had to be multilingual. This was a much greater challenge, but one that would be impossible to avoid. The question was how multilingual access would be provided and in what languages. It was decided to offer the WDL interface in the six official languages of the United Nations: Arabic, Chinese, English, French, Russian, and Spanish. A seventh language, Portuguese, was added, as the National Library of Brazil became a co-founder of the project and a major early contributor of content. For each of these languages, the goal was to provide a uniform user experience, with all navigational information and metadata in all seven of these languages.

It was further decided to offer translation by professional translators and subject matter experts. Machine-assisted translation tools would be used to boost productivity and lower costs, but qualified humans would remain in the loop. Notwithstanding the incredible progress made in recent years in machine translation, the assumption was that only translation by qualified professionals would do justice to the type of content being presented: rare and historic documents, which by their very nature are deeply embedded in national cultural and linguistic contexts. Various other choices in theory might have been made (e.g., use of volunteers on the “wiki” model or resort to “on the fly” machine translation) that might have allowed for a larger number of languages (or freed up resources spent on translation for other tasks, e.g., faster growth in adding content to the WDL), but these approaches were rejected, at least for the moment, as inappropriate for what the WDL was attempting to achieve.

Universality. By definition, the WDL had to be universal, both with regard to participation and content. The WDL had to be open to libraries, archives, and museums from every country in the world: to

any institution that held and was ready to provide important content with bearing on the collective history of humanity. It also had to include material from and *about* all countries and cultures, on the assumption that all had contributed to the heritage of humanity and all had material worthy of inclusion—over a range of time periods and in different formats.

The requirement for universal participation immediately raised the question of capacity. It was obvious that cultural institutions in many countries, particularly but by no means exclusively in the developing world, had little or no capacity to digitize their collections for inclusion in the WDL or in any other project. To aspire to the goal of universal participation, the WDL thus had to include from the beginning a commitment to capacity building and technical assistance—to working with libraries and other partner institutions on acquiring the equipment and skills that would enable these institutions to participate.

The Library of Congress already had experience in the early 2000s in providing equipment, software, and training to libraries in Russia (Moscow, St. Petersburg, and Novosibirsk) and Brazil (Rio de Janeiro) in support of bilateral cooperative digital projects with these countries. Creation of a world digital library with universal or near-universal participation would require replication of these efforts on a vastly wider scale. The Library of Congress lacked the mandate and the resources (financial and human) to take on such a task, but it was hoped that assistance would be provided by the wider international community.

Functionality and user added value. This third set of projected requirements encompassed numerous dimensions and involved a great deal of discussion in the planning, prototyping, and development stages of the project, both within the Library and with prospective WDL partners.

Content selection was arguably the first and most important area in which the WDL needed to add value. Selection in turn was linked to preferences and assumptions about the desired end-state size of the WDL. At one extreme, a WDL could be designed as very small and selective: for example, a “top ten treasure” list of items from each of the 193 UN member countries, or fewer than 2,000 items in all. At the other extreme were examples of vast projects such as Europeana and Google BookSearch, which aimed to gather on a single portal tens of millions of metadata records and/or digitized works.

In the end, it was decided to adopt a middle course: to build a representative body of content relating to the history and culture of all countries, with culture defined in the broad anthropological sense, with a heavy focus on special collections and rare and unique documents. Selection criteria were elaborated by a content selection working group (subsequently transformed into the Standing Committee on Content Selection), which met in Paris in October 2007 and again in Cairo in January 2009, and which endorsed the idea of showing content important “for the history of humanity.”⁴ Particular emphasis was placed on including collections already listed on the UNESCO Memory of the World registry, a pre-existing list drawn up by an established nomination and vetting process, and which the WDL had no need to reinvent.

The second area in which the WDL needed to add value was in the discovery and display of content. This related fundamentally to metadata and, by extension, to the search and browse capabilities that particular sets of metadata would enable. The WDL’s declared emphasis on promoting intercultural understanding implied that it should enable users to access sets of content by which they could, for example, compare the achievements of different civilizations in the same historical time period (e.g., 16th century maps or printing in China, Europe, and the Islamic world), track developments over time in the same country (e.g., Egypt in the Pharaonic, Hellenistic, and Islamic periods), and search and browse objects or topics relevant at different times in different periods (e.g., pyramidal structures built in ancient Egypt, ancient Mesopotamia, and pre-Columbian America).

⁴ <http://project.wdl.org/content/contentguidelines.html>.

This requirement implied a heavy reliance on traditional metadata, much of which would need to be enhanced by adding new fields and values to the bibliographic records provided by contributing libraries. To the extent possible, metadata had to be provided at the item level, with the item defined, as often as feasible, as a single photograph within an album or a single map within an atlas. Only in this way could the individual photograph or print in one album be located and paired with an analogous photograph or print from a different album (or with a book, manuscript, map, and so forth). This approach was certain to be expensive and time-consuming, and it to some extent cut against the broader trend in the library community, in which users and providers of content alike were looking either to open-ended search or to user tagging to replace (or at least supplement) traditional metadata.

Going beyond improved discovery and display of content, it was also decided that the WDL would *explain* and *interpret* content. This was to be done primarily through descriptions, which were made an integral part of the WDL metadata scheme and provided for all items, as well as through curator videos for selected items. Users of many first generation digital library projects had been complaining about what they perceived as libraries' practice of throwing vast amounts of content up on the Internet and leaving it to the users to figure out what it was. So the decision was made to enhance every item-level display record with a paragraph-length description. Detailed instructions on how to write and edit descriptions have since been developed, but in essence the descriptions were and are intended to answer a simple, two-part question: "what is this thing and why does it matter?" The answer to this question was to be given in straightforward, non-technical, jargon-free language, accessible to the interested general public and readily translatable into the seven WDL languages. Many bibliographic records already contained note fields with detailed information about particular rare books, manuscripts, maps and so forth, or such information could be gathered from exhibit catalogs and labels, scholarly journals, and finding aids. However, in many cases this information did not exist and would need to be researched and written from scratch.

These early choices regarding how content would be cataloged and displayed, along with the requirements of multilingualism, largely determined the choice of technical architecture. A distributed system that simply aggregated metadata or provided a federated search was rejected in favor of centrally gathering the content, which then would be distributed worldwide through a content delivery system to maximize speed and performance. All metadata and images would be ingested from partners, and the WDL would use a standard metadata scheme for all items, with values provided for all essential fields and translated into the seven interface languages. This approach was intended to ensure a uniform user experience for all content, regardless of the institution providing the content and the manner in which that content was displayed on the institution's own website (which the WDL always linked back to at the item level in any case). This choice of architecture in turn made possible certain other features, for example text-to-speech conversion for metadata and descriptions and standardized content download options.

In setting these requirements, the planning team did not completely disregard questions of cost, but it treated these questions as something to be considered in a second stage of the analysis. As indicated, the first set of questions revolved around what a WDL should look like and what would make the project really "worth doing." They did not focus on what it might cost to build and sustain the WDL and on what combination of government, private, and foundation sources might be prepared to meet this cost. But considerations of cost were not entirely overlooked. Some rough initial estimates regarding costs were put forward. In the discussion paper prepared for the December 2006 UNESCO Experts Meeting, for example, the figure of \$27,000,000 over a five-year period was floated as a credible estimate.

3. Implementation and Results

Having settled on and defined these three main sets of requirements, the Library of Congress and its international collaborators began to implement the project. There is no need to recount in detail this process, but a few milestones are worth mentioning.

In December 2006, the Library and UNESCO jointly convened an experts meeting at UNESCO headquarters in Paris to solicit input about the proposed project from librarians and technology experts from around the world. The results of the Paris meeting included the establishment of working groups for technical architecture and content selection and a decision that the Library of Congress would develop the prototype of a future world digital library for presentation at the October 2007 UNESCO General Conference. The prototype was presented as planned, with content provided by six partner institutions: the National Library of Brazil, the Bibliotheca Alexandrina of Alexandria, Egypt, the National Library and Archives of Egypt, the National Library of Russia, the Russian State Library, and the Library of Congress.

There were also breakthroughs in technical assistance and capacity building. As a pilot project to gain experience and help in better understanding the challenges and costs of providing such assistance, in 2006 the Library of Congress concluded an agreement with the National Library and Archives of Egypt to provide high-end equipment to NLAE for the purpose of digitizing Arabic scientific manuscripts for inclusion in the WDL. Similar agreements were concluded with and similar sets of equipment provided to the Iraqi National Library and Archives and to the National Library of Uganda (the latter with funding from Carnegie Corporation of New York).

Following eighteen months of intensive planning and development, www.wdl.org was officially launched at UNESCO headquarters in Paris on April 21, 2009. Developed at the Library on the basis of the 2007 prototype, the site featured content contributed by institutions from eighteen countries, including the national libraries of China, France, Israel, Japan, Russia, Serbia, and Sweden as well as major university libraries from several nations.⁵ The national libraries of Egypt, Iraq, and Uganda were all “founding partners,” and some content scanned at their newly-established WDL digitization centers was included in the launch version of the site.

Following the launch, developing a governance structure for the WDL became a priority. To get the project off the ground, in 2006-2009 the Library of Congress concluded a large number of agreements with WDL partner institutions regarding the use of their digital images and metadata on www.wdl.org. These agreements differed with respect to the wording of key provisions, duration, the rights and responsibilities of the parties, the language of the agreement, and so forth, with variations being the product of the different legal offices involved in the drafting of such agreements. In 2010 the WDL was restructured to become a loose multilateral institution with a uniform set of rules, a permanent governance structure, and a simple statement of rights and responsibilities, including intellectual property rights. These provisions were codified in a charter that all partners were to sign and that was made equally authoritative in the seven interface languages. Adopted in March 2010, the charter provides for an annual partner meeting, an Executive Council elected by the partners, and permanent committees for Technical Architecture, Content Selection, and Translation and Language. The charter also provides for an institutional project manager, responsible for maintaining and building the WDL website, and designates

⁵ “Build It, and They Will Come,” *Library of Congress Information Bulletin*, May 2009, <http://www.loc.gov/loc/lcib/0905/wdltech.html>

the Library of Congress as the institutional Project Manager for the period 2010-2015.⁶ Partners join the WDL by acceding to the charter, in effect concluding an agreement not just with the Library of Congress, but with all of the WDL partners.

Since the early startup phase, the WDL has continued to grow, operating under the terms of the charter. For the most part it has continued along the lines set in the initial planning stage, with some inevitable adjustments in response to changes in technology, user feedback, and partner preferences. Progress has been made on all fronts, and the basic decisions made in the planning stage appear to have been validated, in the main if not in all particulars.

The commitment to multilingualism has paid off. Content on the WDL is in 86 languages, with the most heavily represented languages being Spanish, English, Arabic, German, Russian, French, Chinese, Latin, Japanese, and Portuguese (ranked by number of items; the number of pages/images would yield a different result). More can and will be done in this area, and the WDL especially welcomes content contributions in lesser known and endangered languages.

Providing multilingual access also has been a success. Each of the seven interfaces is extensively used, with the Spanish-language site by far the highest. In 2009, the Spanish-language interface accounted for 33.1% of all pages viewed, followed closely by English (30.6%), French (11.3%), Russian (8.8%), Portuguese (7.7%), Chinese (6.5%), and Arabic (2.0%). By 2011, Spanish had surged even further into the lead, with the Spanish interface accounting for 58.4% of pages viewed, followed by English (17.1%), Portuguese (14.4%), Russian (4.0%), Chinese (2.9%), French (2.5%), and Arabic (1.4%). A key task for the future will be to sustain and continue building on the high levels of usage in Latin America and the Iberian Peninsula, in both Spanish and Portuguese, while increasing usage of the other languages, Chinese and Arabic in particular.

The commitment to universality remains valid. As of this writing (mid-September 2012), the WDL has 159 partners from 75 countries, in all continents and UNESCO regional groups, in both the developed and developing world. This is still a long way from universal participation, but it clearly represents progress in that direction.

Universal content coverage remains a value and is reflected in content selection priorities and the setting of production schedules. The WDL website contains 6,330 library items comprising a total of nearly 300,000 images, contributed by 84 institutions in 42 countries. Some content about each UN member country is included and has been since the 2009 launch. Significant amounts of content from numerous partners are at various stages of the production pipeline, and new partners from additional countries are in the process of joining the project. Comprehensive coverage of all countries is still a long way off, but the project has set a goal of having a minimum of 100,000 rare and unique items on the WDL to provide such coverage.

High levels of usage and user satisfaction suggest that the decisions regarding functionality and user value-added have paid off. On its first day, the WDL received over 600,000 visitors from every country in the world. Since then, more than 21 million people have visited the site, accounting for some 135 million page views. Various web awards have been won, and thousands of user comments have been received, in each of the seven WDL languages, for the most part overwhelmingly positive. Nearly 5 million other sites link to the WDL, testifying to the enthusiasm with which the project has been embraced and ensuring high rankings on Google and other search engines. Total users averaged about

⁶ Information about organization and governance – including the charter in the seven WDL languages -- can be found on the WDL project website at <http://project.wdl.org>.

420,000 per month in 2011, with the highest numbers by country from Spain, Mexico, Brazil, the United States, Argentina, Chile, Colombia, Portugal, the United Kingdom, Russia, and France.

The decisions to standardize metadata and provide item-level descriptions have contributed to the appeal of the project with users, and the choice of technical architecture, originally quite controversial, seems to have been vindicated. Even at its current, still relatively modest size, a search of the WDL for the term “Islam” yields 365 books, manuscripts, prints, newspapers and maps, provided by 21 institutions in 16 countries.⁷ The architecture and metadata enable the user to quickly browse through all 365 items, narrow the search by place of origin, date, language of the document, format, other or additional subject (beyond Islam), and contributing institution. The user can instantly see the full item, zoom in to high levels of detail, access all of the metadata in six additional languages, listen to the metadata using voice to speech conversion, download the content in PDF form, and learn a lot more about the item from the description, written by a qualified expert. These features would be impossible to provide in a distributed system, particularly one involving links to partner institutions in parts of the world still struggling with poor connectivity and low bandwidth.

Sustaining high levels of quality and added value for users will be an ongoing challenge for the WDL. So far, however, there has been no deterioration—indeed the trend has been in the other direction, as quality and standards have consistently risen as the project gains experience and benefits from new contributions from new partners. The launch version of the WDL was developed with a tiny staff and against enormous time pressure. Since that time, the quality of translations has been upgraded by the recruitment of dedicated staff, deployment of new translation management tools, and the putting into place of mechanisms to respond to comments and feedback from users, who are often quick to detect errors and omissions. The quality of the metadata also has been improved, through the recruitment of a full-time professional staff, the development of a new metadata application, continued refinement of standards, and integration of the Virtual International Authority File and other tools into the cataloging process.

The quality of the descriptions has been steadily upgraded. Partners providing content to the WDL have been given better guidance and now have more time to work with their curators to explain their content in appropriate depth and detail, with time and procedures available for back-and-forth between the WDL production team and the content providers to check facts and clarify ambiguities. In the startup phase of the process, many descriptions were written quickly or cobbled together from pre-existing sources. Now, the annotation of certain collections at the item level has become a much larger and better organized operation, encompassing within it several major subprojects that are producing high quality primary research by recognized experts.⁸

Most importantly, the overall quality level of the content, in terms of rarity and cultural and historical importance, continues to rise, as new partners join and contribute marvelous items from their collections. The launch version of the WDL contained a fair share of top treasure content –*Miroslav's*

⁷ Bosnia, Brazil, Egypt, Germany, India, Iraq, Kazakhstan, Lebanon, Mali, Morocco, the Netherlands, Pakistan, Qatar, Saudi Arabia, Slovakia, and the United States.

⁸ Examples include the item-level descriptions for the photographs in two major collections held by the Library of Congress that are on or being added to the WDL: the 1871-72 *Turkestan Album* survey of Central Asia and the *Prokudin-Gorskii* collection of early photographs of the Russian Empire, by Professor William Brumfield of Tulane; annotations of Arabic and Persian scientific manuscripts by academics from Columbia and Cambridge universities; and the annotation of rare Chinese books and manuscripts by retired Chinese-language experts from the Library of Congress.

Gospel from the National Library of Serbia, the famous “Devil’s Bible” from the National Library of Sweden, anthologies of very old and rare Asian materials from the National Library of China and the National Diet Library of Japan, and similarly distinguished content from other partners. But many of the approximately 1,200 items initially in the WDL were included to provide minimal coverage of each of the 193 UN member countries. The latter items were carefully chosen and of generally high quality—mostly maps, photographs, or 18th and 19th century books—but they were not in most cases “top treasures.” This material is now being supplemented by a steady wave of rare treasures from such great institutions as the Bavarian State Library, the national libraries of France and Spain, the Laurentian Library in Florence, the Estense Library in Modena, and many others. It will take some time before all countries, particularly those in the developing world, are represented by the rarest and most important documents from and about those countries, but the trend is in this direction.

The level of quality of the content in the site also has benefited from the designation, endorsed in the 2011 WDL business plan and discussed at the 2011 partner meeting in Munich, of several areas in which the WDL will develop concentrations of important content, working proactively with partners and prospective partners. These areas include Mesoamerican codices, Chinese rare books and manuscripts, Arabic scientific manuscripts, early photographic surveys of empires, and treasures from medieval and Renaissance Europe. The WDL has made great progress in all of these areas, and the long-term potential of this approach can be seen most clearly in the case of the codices, where documents from the 11th to the 16th centuries relating to the history of the Aztec and Mayan peoples have been contributed, so far, by libraries, archives, and museums in Mexico, Spain, Italy, Germany, Sweden, and the United States.⁹

4. Lessons and Implications

So what have been the lessons learned? What about the WDL has worked, and what has fallen short of expectations? And what implications might the WDL experience have for other digital library projects and for the overall theme, “Beyond Access: Digitization to Preserve Culture”? A few conclusions can be drawn.

Multilingualism has been a big success for the WDL, with users if not necessarily with international funding sources. Heavy usage of the Spanish, Portuguese, and other interfaces suggests that, while the volume of content on the Internet in languages other than English has grown astronomically in recent years, there remains a demand for sites that provide high-quality cultural and historical content in a range of languages. International library projects can draw from the WDL the lesson that investment in providing multiple-language interfaces pays off in increased access. But multilingualism is an expensive proposition, and human and financial resources will be needed to encourage and sustain the development of multilingual digital library projects.

The capacity building mission of the WDL has been at best a very mixed success. Staff at the Library of Congress and at partner institutions in Cairo, Baghdad, and Kampala made heroic efforts in 2006-2009 to install equipment and train staff and to scan initial batches of content for inclusion in the launch version of the WDL. Production has continued at these three centers, and various other capacity building activities—training workshops, visits, and the provision of software tools—have taken place. But

⁹ For documentation regarding this focal area, see Mesoamerican Codices Meeting, May 19-21, 2010, Mexico City, Mexico, http://project.wdl.org/content/mexican_codices/index.html.

generally speaking there has not been much interest in or many resources available for expanding digitization capacity in developing countries in connection with this project.

The WDL has called for interested institutions and organizations—WDL, IFLA, UNESCO, and perhaps others—to work together on a comprehensive needs assessment of digitization in developing country institutions and on the development of low-cost solutions and sources of support for implementation. But so far little has happened in this regard, even as documents continue to be destroyed by war and natural disaster in many parts of the world. Capacity building remains, like multilingualism, part of the WDL's core mission and efforts along these lines will continue, but the volume of resources committed by the international community devoted to work in this area has been disappointingly limited.

Judged by user responses and the enthusiastic participation of so many great libraries, the WDL's effort to provide a high level of intellectual and functional added value—metadata, descriptions, translations, and other features—also seems to have been validated and would appear to be worthy of emulation by other projects. Already in 2006 it was becoming clear, and it is even more apparent in 2012, that non-profit cultural institutions cannot compete on quantity with large commercial firms, and especially not with commercial firms able to access vast amounts of free, user-generated content which these firms can then monetize in one way or another. They can compete qualitatively, by presenting and interpreting the treasures in their vast holdings and drawing upon the intellectual capital embodied in their metadata, finding aids, exhibit catalogs, collection guides, and the minds of their current body of curators.

This is the path that the WDL and some other projects have chosen, and one that would seem worthy of emulation. At the very least, it is to be hoped that cooperative international projects such as the WDL will continue to remind people that cultural heritage—what the organizers of this conference have chosen to call the “memory of the world”—must be primarily about content, about the cultural artefacts and information that carry the collective memories of countries and peoples. Whether this will happen remains to be seen. Cultural institutions are under financial strain and trying to focus their available resources. As they do so, many appear to be downplaying their traditional strengths in curatorial and subject matter expertise in favor of a concentration on information science.

Resources are a serious issue. As noted, the WDL did not start out asking the question: how can we use our available funding to create a World Digital Library. Rather, the question was, what should a WDL be and do? The question then became what will it cost and can the needed resources be found to support the effort. This to a great extent remains a question that members of the international community need to debate and resolve, not just with regard to this one project, but with regard to the overall issue of “Memory of the World in the Digital Age.”

Strategies for Building Digital Repositories

A Persistent Digital Collections Strategy for UBC Library

Bronwen Sprout and Sarah Romkey

Abstract

In early 2011, UBC Library began work on creating a digital preservation strategy in collaboration with Vancouver-based Artefactual Systems. UBC Library and Artefactual ran a number of pilot projects to ensure that the strategy is pragmatic, tested, and involves proven technical solutions and business processes that work in our environment and towards our goals. The persistent digital collections strategy developed for UBC Library consists of using the open-source Archivematica digital preservation system to provide preservation functionality for the Library's digitized and born-digital holdings. The strategy identifies the software requirements, existing and new system components, staffing and business processes that can be implemented to establish operational digital preservation systems and processes by the end of 2012. The paper will discuss the strategy generally and cover three areas of implementation in greater detail: UBC Library's Rare Books and Special Collections, cIRcle, our DSpace-based institutional repository, and CONTENTdm, UBC Library's access system for digitized objects.

Authors

Bronwen Sprout is the Digital Initiatives Coordinator at UBC Library. She coordinates the development and management of the Library's locally created digital collections, including digital preservation activities. She has over ten years' experience working on digitization projects and managing digital collections. Prior to joining UBC Library in 2005 as Digital Initiatives Librarian, Bronwen worked as a cybrarian at a dot-com, and as a consultant for small businesses and non-profits, assessing information needs and implementing a variety of web-based solutions.

Sarah Romkey graduated from University of British Columbia's joint Master's in Archival Studies and Master's in Library and Information Studies program in 2008. She has since that time worked in UBC Library's Rare Books and Special Collections division initially as the Librarian and Archivist for the Chung Collection and since 2009 as the Rare Books and Special Collections Archivist. Sarah is responsible in this role for acquiring, providing access to and preserving the archives of individuals and organizations external to UBC.

The University of British Columbia (UBC) Library's digital preservation strategy supports the Library's strategic focus on managing collections in a digital context and reinforces the University's commitment to make UBC research accessible in open access repositories. In 2010, the Library's newly launched strategic plan guided the development of a Digital Initiatives unit and related decisions to hire a Director of Digital Initiatives and to outfit and staff a digitization centre. The Digital Initiatives unit is a key part of the Library's effort to adapt to the evolving needs of faculty and students and to support teaching, research and learning at UBC. The unit's goal is to create sustainable, world-class programs and processes to make the collections and research at UBC available to the world and to ensure the authentic, long-term preservation of these digital holdings for the future. As the new unit became established, one of the obvious functional gaps was a preservation system for digital master files. As we began work on a digital preservation strategy however, it quickly became apparent that there were crucial digital preservation gaps with other Library-created and/or managed material as well.

This paper will address the pilot and early stages of the implementation of UBC Library's digital preservation strategy. The Library has been working on this strategy with Artefactual Systems Inc., the Vancouver-area company behind the open source Archivematica and ICA-AtoM software, since early 2011. The first year of the project was devoted to developing and testing the preservation strategy and included installation and testing of the Library's Archivematica instance and the development of copying workflows and procedures. The second and current year of the project is the implementation phase; we hope to be able to offer preservation services by the end of 2012.

Archivematica is one of the key components of our persistent digital collections strategy. The software "uses a micro services approach to provide an integrated suite of free and open-source tools that allows users to process digital objects from ingest to access in compliance with the OAIS model."¹ Archivematica will provide preservation functionality for the Library's digitized and born-digital material, managing functions and activities such as checksums, virus checking, preservation metadata, and file format normalization. It will interface with our existing systems (DSpace, CONTENTdm, ICA-AtoM) and we will use existing Library staff resources to integrate digital preservation activities into our workflows. Although Archivematica is a key component, our preservation strategy is not solely focused on this software. The strategy also includes software requirements, existing and new system components, staffing and business processes.

Artefactual's initial task in establishing our strategy was to carry out was a gap analysis. The gaps in the Library's digital preservation activities were largely identified using the ISO-OAIS reference model.² The consultants looked at the digital material we were creating or managing locally and, referencing OAIS, asked: "what is not being adequately preserved?" Following that, several areas were identified for pilot projects, including:

1. Rare Books and Special Collections and University Archives – both pilot projects involve born digital records and are working with legacy and newly acquired material
2. cIRcle, our DSpace-based institutional repository – contains more than 40,000 items, mainly theses and dissertations but also a variety of research and teaching materials
3. Digitization projects in Digital Initiatives – CONTENTdm is UBC Library's access system for digitized objects and currently contains about 80 collections and over 200,000 images

Once the gaps were identified, a project team was struck. The team includes Artefactual staff as well as representatives from the areas of the Library identified and/or affected by the pilot projects: Digital Initiatives, Library Systems and Technology, University Archives, and Rare Books and Special Collections. A student assistant from the School of Library, Archival and Information Studies was also a part of the project team and assisted hugely with the testing. The team meets monthly and UBC Library members work closely with Artefactual to test the software and iterate the development of workflows that will integrate with our local practices.

One of the major outcomes of our digital preservation program will be the ability for Rare Books and Special Collections (RBSC) to archive born digital material. The mandate of this division of the UBC Library includes the collection of archival material from private creators about the various aspects of life and industry in British Columbia. As of 2012 the division has in its holdings almost 700 archival fonds

¹ "What is Archivematica," accessed August 16, 2012, https://www.archivematica.org/wiki/Main_Page.

² "Open Archival Information System," accessed August 16, 2012, http://en.wikipedia.org/wiki/Open_Archival_Information_System.

created by a wide variety of people and organizations, dating from the gold rush era to the present day. Since the time when digital technology became ubiquitous in homes and offices, RBSC has acquired born-digital records on external media (primarily floppy disks, CD ROMs and DVDs) only in the context of the media arriving unexpectedly in primarily analogue acquisitions. While this media was presumed to be of archival value and kept with the fonds, one could say that the digital media was not processed in the sense that no attempt to preserve the digital records separately from the physical media was made.

The University Archives division of UBC Library, responsible primarily for university and university-related records, had similarly begun to receive digital media, or requests to acquire digital media, from university departments. In 2011, the archivists from both divisions had the opportunity to participate in UBC Library's persistent digital collections strategy pilot project. The aim for these two divisions in the pilot project has been to develop a method for both processing legacy media in existing collections, and acquiring and processing born-digital material from new or ongoing collections.

After identifying collections in RBSC for which legacy digital media existed, the project team prepared to test the preservation of the files. Appropriate hardware was procured, including a 3.5 inch floppy disk drive and, with more difficulty, a 5.25 inch floppy disk drive. A basic workflow was developed which involved creating a Submission Information Package (SIP) from the files on the external media and feeding it through Archivematica, which would create an Archival Information Package (AIP) and a Dissemination Information Package (DIP) for access in ICA-AtoM archival description software.

Unsurprisingly, technological challenges have abounded during testing of external media, particularly in regards to floppy disks. A number of the disks tested were found to be unreadable, and possibly have been unreadable since their initial accession to the archives. The 5.25 inch floppy drive has proven to be particularly challenging to operate. A number of files which could be copied from the disks have been in outdated formats for which the proprietary software no longer exists. Some optical disks were found to be created from proprietary sources (such as photo finishing stores) which also required proprietary software to open or copy the files, even if the files themselves were in commonly used formats. At this point, we have sufficient experience processing external media ourselves to feel comfortable approaching an external company to take on some of the processing. Our pilot testing has allowed us opportunity to create documentation for several different use cases, including transferring media from optical disks, 3.5 and 5.25 inch floppy drives, and external drives, which we can use to guide future work as well as our negotiations with vendors.

Processing legacy digital media has also raised a number of intellectual problems. One is the arrangement of the digital media into series or files within the collection. Although archival theory and practice tells us that the physical container of records may or may not reflect the intellectual arrangement of the records, the processing archivists who had accessioned these records were limited in their options for arranging the media into series, being unable to separate the records from the media itself. Thus it is not uncommon in our finding aids to find series titled, "Electronic media" or similar. Actual reading of the records through a file viewer often reveals evidence of original order (the principle by which an archivist would ordinarily arrange records into series and files) such as folders and naming conventions applied by the creator. Furthermore, it may appear that records on the digital media share "archival bond"³ with analogue records in the same collection by virtue of their association with the same activity

³ Archival bond is defined as "The network of relationships that each record has with the records belonging in the same archival aggregation." From "The InterPARES Project Terminology Database," accessed August 23, 2012, http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=a&term=6.

of transaction of the creator. Does one re-arrange the collection into series based on this newly discovered evidence, or maintain the arrangement provided by the first processing archivist? At this juncture, decisions are being made on a case-by-case basis.

It is one of an archivist's core duties to *select* records for permanent retention within a collection. Similar to the challenge of arrangement described above, the processing archivists of the past did not have an effective way of selecting the records contained on digital media, lacking a method for opening the files and inspecting their content. They were therefore only able to use what metadata existed on the label or container for the media and make assumptions about the content. While some disks have yielded archival documents such as book manuscripts and digital photographs of which no other copies are known to be extant, other disks have yielded routine administrative records that would not have normally met an archivist's criteria for retention, lacking sufficient evidential or informational value to be of use to researchers.

Finally, an examination of the rights transferred with the digital material must occur in order to determine the institution's ability to preserve and provide access to the material. While UBC Library's standard donation form for archival material transfers copyright to the institution, understandably not all archival creators are comfortable with such an arrangement and so they are accommodated by transferring the legal ownership of the physical material while maintaining their intellectual property rights. Because this legacy digital media was acquired in a primarily analogue context, no licensing was written into the agreements to allow online access or the creation/migration of digital copies for preservation reasons.

While testing was still occurring with the legacy media from RBSC and University Archives collections, it seemed that sufficient lessons had been learned to pursue what could be considered our "first" born-digital archival acquisition; first in the sense that we proactively sought a primarily born-digital acquisition and would therefore have greater control over many aspects of the acquisition process. RBSC had been approached by a photo-journalist who desired to donate his own born-digital photographs from the past seven years. The acquisition was from our perspective a desirable test case because first, the subject matter of the photographs (fishing boats and fishing practices) complements other collections in our holdings, and second, the donor has been able and willing to give answers to our questions regarding the creation of these images.

The information gathering stage was important in order to address the same issues we had faced with our legacy media project. A questionnaire for a donor interview was developed, adapted from *Born Digital Collections: An Inter-Institutional Model for Stewardship (AIMS) survey for Personal Digital Archives*⁴ and the *Paradigm records survey*⁵ published by the Bodleian Library. While the AIMS and Paradigm surveys provided ample opportunity to ask the donor about his technological and organizational processes, additional questions were added to address questions of rights and access: did the donor wish to retain his copyright? If so, what measures was he willing to allow UBC Library to take in order to preserve and provide access to his born digital records? Were there any rights attached to the records that he did not have the ability to transfer or license to the institution? (See Appendix A for our survey instrument).

The donor's answers to the questionnaire allowed us to prepare both technologically and intellectually for the transfer of the material. From a technological point of view, the feasibility and

⁴ "AIMS Born Digital Collections: an Inter-Institutional Model for Stewardship," accessed August 31, 2012, <http://www2.lib.virginia.edu/aims/>.

⁵ "Welcome to paradigm," accessed August 31, 2012 <http://www.paradigm.ac.u/>.

appropriateness of using digital forensics tools and processes was explored. The donor wished to transfer the records on an external hard drive, which he had purchased explicitly for the purpose of donating his digital records. Initially we felt that using a write blocker, a piece of hardware used to prevent the inadvertent alteration or damage of files and documents when external devices are used to transfer files, would be an appropriate safe-guard. However, it became necessary to delete files from the external drive before the transfer could take place due to an accidental inclusion of files not intended for donation, and so the use of a write blocker became an inappropriate step.⁶ This step was also deemed not *as* necessary as perhaps would be appropriate in other use cases: in the future, we anticipate that entire working hard drives (as opposed to copies of digital records saved to external media) will be imaged in order to preserve the records. In that use case a write blocker would be necessary to ensure the records, being the original records created by the donor, will be imaged authentically.

Based on our experience processing legacy media and the external drive we began to perceive there to be different *levels* of digital preservation that might be appropriate to guide future acquisitions:

Level of Preservation	Use case	Minimum tools
Level 1	Imaging files from working hard drives, directly from creator	Write blocker, file synchronization software (e.g., Grsync)
Level 2	Imaging/copying files from external media	File synchronization software (e.g., Grsync), write blocker if possible
Level 3	Preserving external media only	Hardware necessary to open media (floppy drive, optical drive, etc)

“Level 1” preservation would entail imaging of working hard drives, when a creator is prepared for the archivists to examine all of their digital records and image files directly from their working computer. Digital forensics tools such as write blockers would be necessary for this level of preservation to ensure the authenticity of the original files is maintained. “Level 2” preservation would entail making authentic copies from external media such as floppy or optical disks and external hard drives. Software tools such as Grsync⁷ and functions such as checksums will be used to ensure the copied files are authentic copies of the files provided on the external media. Ideally, a write blocker would be used, but would not be considered as necessary as in Level 1. “Level 3” preservation may be a last resort to maintain bit-level records on external media that cannot be copied authentically due to technological problems. We have developed use case documentation for acquisitions made through external media which include the use of a write blocker, even though it has not been used for this particular pilot test.

⁶ Using a write blocker and making a forensic image of the hard drive contents would have inevitably meant including the files which the donor requested us to delete. We decided that this was not acceptable from a donor-relations point of view.

⁷ “Grsync: Home,” accessed August 31, 2012, <http://www.opbyte.it/grsync/>.

To ensure that we have the necessary rights or licenses to preserve and provide access to the donor's material, we developed a donor agreement form specific to the issues presented by born-digital material. It was our initial hope that our existing donation agreement form would suffice with some alteration. Our realization though was that born-digital acquisitions go beyond the realm of physical, and even intellectual, transfer of material in the analogue world; what is really needed is the ability to gain a license for the rights needed to make copies of records for preservation purposes, migrate the records to new formats now and in the future, and provide access to users either on the internet or through a computer terminal in our reading room. Once again, AIMS and Paradigm documentation provided models, which were combined and modified with additions from a non-exclusive digitization and distribution agreement developed by UBC Library for managing rights associated with digitization projects.⁸

At the time of writing, we are at the stage of using Grsync to copy files from the donor's external hard drive, with the intention of running the acquisition through Archivematica in the fall of 2012. One question that remains to be explored in this pilot project is the extraction of creator-embedded metadata for use in the access system. The creator of the records meticulously adds geographical and subject indexing terms as well as brief descriptions to many of his photographs, which would be helpful for access purposes if they can be re-purposed.

Whereas digital acquisitions are still a fairly minor aspect of RBSC and Archives' holdings, cIRcle, UBC's institutional repository, is solely concerned with digital material. And whereas processing legacy digital media raised a number of intellectual questions, developing the Archivematica integration with DSpace and CONTENTdm has so far raised mainly technical issues.

cIRcle is a DSpace-based open access repository which currently contains over 40,000 items, about half of which are theses and dissertations. In addition to the comprehensive retrospective digitization of theses dating back to 1919, UBC has offered optional online submission of electronic theses and dissertations (ETDs) since 2008 and mandatory online submission for the last year. Thus, the ETDs in cIRcle are considered the University's official copies and were a major driver for the development of a digital preservation strategy for the repository. Because we use a distributed submission model to add new content to the repository, in terms of our digital preservation plan, DSpace will continue to function as the submission and access tool and full preservation management will take place separately in Archivematica. Archivematica will accept submissions from DSpace and preserve them without creating access versions. Submission to Archivematica will take place when submission is made to DSpace, without affecting the user interface. The identifier field (handle) will link the DSpace access record and the Archivematica preservation version. Archivematica 0.8 was designed to recognize DSpace exports and parse them into the Archivematica preservation version.⁹

Though it is not uncommon to refer to DSpace as a preservation system, the software does not have full preservation functionality. Archivematica performs many preservation actions on the DSpace files: it unpackages them, verifies their checksums, assigns unique universal identifiers, checks for viruses, identifies and validates formats, extracts technical metadata and normalizes the files to preservation formats. All metadata are captured as fully standards compliant PREMIS –METS. At this point, a couple

⁸ Readers may also be interested in consulting *Research Library Issues (RLI) no. 279*, recently published by the Association of Research Libraries, which discusses digitization of analogue collections and provides a model gift agreement for mixed intellectual property rights (<http://www.arl.org/news/pr/rli279-7aug12.shtml>).

⁹ "DSpace exports," accessed August 16, 2012, https://www.archivematica.org/wiki/DSpace_exports.

of items regarding the DSpace integration are incomplete; our next steps include a retrospective migration of DSpace content into Archivematica and the development of an automated export/ingest trigger for newly-submitted items (so that items approved for submission in cIRcle are automatically exported to Archivematica).

A companion access system to cIRcle, CONTENTdm is UBC Library's access system for locally digitized objects. The digital collections available in our instance of CONTENTdm date back over 10 years to projects that were created long before the development of the new Digital Initiatives unit or the consideration of digital preservation. Collectively they document a diverse range of people and places, activities and events, and serve as a resource for students, historians, genealogists, and other researchers. The content of these collections is largely drawn from RBSC and Archives and ranges from rare maps and books to historical newspapers to images from British Columbia's forestry and fishing industries. A large part of the collection consists of images which present a visual record of UBC's growth and development over the past century. Like most sites using CONTENTdm, we digitize analogue materials, store the master versions of the digitized objects on network drives and upload the access copies, along with descriptive metadata, to CONTENTdm for public access. In our case, the master files were not adequately linked to our access copies (a filename based on the location of the physical item loosely links the two versions, but is prone to human error). And although we do participate in a Private LOCKSS Network (PLN) that has the capacity to store our CONTENTdm collections, it will not preserve master copies because we do not make them available online for LOCKSS to crawl. Therefore, one of our pilot projects includes the development of an API to link Archivematica and CONTENTdm.

We plan to ingest master digitized objects into Archivematica which will generate the access version for automatic upload into CONTENTdm. The SIP is ingested into Archivematica and when it has moved through to become an AIP the user is given the option of using CONTENTdm to make the DIP available to end users. Two methods for submitting the item to CONTENTdm are available: a one-click import into CONTENTdm as well as the option to save the item as a set of files suitable for importing into CONTENTdm using the Project Client (so Optical Character Recognition can be applied using the Project Client, etc.). Work is currently underway on a bulk DIP upload with metadata attached to each object via a user-supplied CSV file. With either method the regular CONTENTdm approval/indexing needs to be applied before the item appears to end users. Although the option exists to add metadata in Archivematica, most of the descriptive metadata for uploaded objects will be created and edited in CONTENTdm because it better facilitates batch processing. (Descriptive metadata will be added to the AIP by syncing the DIP in CONTENTdm and the AIP in Archivematica.)¹⁰ This will change our workflow considerably as all material will have to go through Archivematica to be added to CONTENTdm. However, we plan to add this step to existing workflows and build the work into existing job descriptions rather than hiring or designating specific staff to take on the role of digital preservation.

Mid-way through 2012, many implementation tasks remain. Archivematica testing is ongoing as are the DSpace and CONTENTdm integrations. A large retrospective project to add DSpace and CONTENTdm content to Archivematica is necessary. We are interested in undertaking a pilot with the Council of Prairie and Pacific University Libraries (COPPUL) PLN we participate in, in which objects ingested into Archivematica would be uploaded to the PLN for distributed, geo-remote storage. We also plan to undertake a research data pilot project using Archivematica to ingest and preserve one or more

¹⁰ "CONTENTdm integration," accessed August 16, 2012, http://archivematica.org/wiki/index.php?title=CONTENTdm_integration.

datasets. We are excited about the potential development of a unified discovery interface for UBC Library digital collections based on the contents of our Archivematica index. Additional effort is needed on our website archiving plan.¹¹ And finally towards the end of 2012 we plan to work with Artefactual to undertake a self-audit based on the ISO Standard for Trustworthy Digital Repositories (an international standard for auditing compliance with the OAIS reference model).¹²

Despite all of the work underway and still to come, we feel we have made tremendous progress over the past year and a half towards implementing a comprehensive digital preservation program. We have a solid plan for our locally digitized material and for the content of our institutional repository in terms of the Archivematica integration with our two primary access systems. Equally as important, we have a greater comfort with and understanding of the challenges around archiving born digital material as the archivists at RBSC and University Archives are anticipating that most, if not all acquisitions in the future will include a born-digital component. Currently, most newly acquired archival collections to these divisions are still primarily analogue, but frequently include external digital media that the donor either created in their normal business or creative practices, or made purposefully with transfer to the archives in mind. This pilot project has helped to solidify what the normal practices and procedures will be, so that digital acquisitions can be more or less systematic, and fewer decisions made on a case-by-case basis.

During the time of our contract with Artefactual, the software was still in alpha. This meant that UBC Library had the opportunity to contribute to the development of the full release (expected for early 2013). We are greatly anticipating this release and are hopeful that our contributions to the development of Archivematica will benefit other institutions as well.

¹¹ Part of our digital preservation strategy is the ability to capture and preserve websites (mainly UBC sites but possibly external sites as well). We tested the open source tool Heritrix, which was developed by the Internet Archive. The software crawls selected sites and stores them as Arc or WARC files. Heritrix comes with a web-based admin module and sites are rendered by Wayback Machine. We found it fairly easy to use but it does require staff time to configure. A similar service called Archive-It is also available from the Internet Archive which includes hosting and remote storage, and this is likely the option we will implement. A major task still outstanding is to assess our library and university websites and decide on the parameters for the crawl.

¹² "TRAC/TDR Checklist," accessed August 16, 2012, <http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0>.

Appendix A: Donor Survey Instrument

Adapted by the City of Vancouver Archives from *Born Digital Collections: An Inter-Institutional Model for Stewardship (AIMS) survey for Personal Digital Archives* and the Paradigm records survey published by the Bodleian Library and John Rylands University Library, and further adapted by UBC Library.

The survey is designed to be used by the archivist(s) during a face-to-face or phone interview with a donor/creator of born-digital records being considered for acquisition.

Donor name:

Donor affiliation (if any):

Archivist name:

1. Digital Material Creation

- 1.1. Are you the only person responsible for creating your digital files?
 - 1.1.1. If not, who else is involved and what is their role? (for each one list last, first, role [author, editor, secretary, proofreader, admin, ...])
- 1.2. Do you maintain digital files created by others?
 - 1.2.1. If yes, how do you separate your files from files created by others?
- 1.3. Do you separate your personal files from your work files?
- 1.4. What is the earliest creation date (roughly) of your digital files?
- 1.5. What is the latest creation date (roughly) of your digital files?
- 1.6. What software and computer system was used to create these files? Did this change over time?
- 1.7. What kind of camera did you use to shoot these images? Did this change over time?

2. Varieties of Digital Material

- 2.1. What kinds of materials are you donating today? Are they in specific formats (TIFF, JPEG)?
- 2.2. Do you create files in both digital and paper formats?
 - 2.2.1. If yes, which files or file types?
 - 2.2.2. Do you also have prints of your photographs that you will donate?
- 2.3. What do you consider the first, best copy of the file(s)? (this could differ by content type)
- 2.4. Roughly how much of each type exist? (in MB, GB, etc)
- 2.5. Do you distinguish in the files or the filing system between published, non published and re published photographs?
- 2.6. Do you think you will want to reuse or republish the photographs that you donate?
- 2.7. Do you have signed rights or clearances from subjects in your photographs?

3. Digital Material Organization

- 3.1. How are digital files named?
- 3.2. Is some kind of version control used?
 - 3.2.1. If yes, list examples
- 3.3. How are digital files organized? Can you give a brief summary about the organization of your digital files? (obtain classification scheme if available)
 - 3.3.1. Are digital files destroyed in regular intervals?
 - 3.3.1.1. What is the interval? (obtain retention schedule if available)
- 3.4. Do you use more than one computer? (e.g. office desktop, office laptop, home desktop, etc)
 - 3.4.1. If yes, how do you synchronize files between different computers? (server, etc)

- 3.5. Please explain fully your process of adding or creating metadata for your photographs.
 - 3.5.1. Which programs do you use to create metadata in them?
 - 3.5.2. Do you use templates?
 - 3.5.2.1. If so, do you use them for one group, all, some?
 - 3.5.2.2. Is there one template or many?
 - 3.5.2.3. Can we get a printout of your templates?
 - 3.5.3. Do you use the same metadata for all the photos, or do you customize it for all, some, etc.?
- 3.6. Do you wish to retain the copyright for the material you are donating?
 - 3.6.1. What rights are you able/willing to assign to UBC Library (access, migration and copying for preservation purposes, etc)
 - 3.6.2. Are there any photographs for which the copyright has been assigned to someone else (e.g. a magazine, etc).

4. Digital Photographs

- 4.1. Which do you consider the original (jpg, raw, tif...?)
- 4.2. Do you use geocoding?

5. Mobile devices and tablets

- 5.1. Do you use computing devices besides your desktop computer to create this material? (e.g. Blackberries, iPhone, iPad or other tablet, Android phone, etc.).
- 5.2. Do you store photographs on these other devices?

6. Email

- 6.1. Do you have multiple email accounts?
- 6.2. Which email program(s) / service(s) are you using? (e.g. Email program provided by your work place, Outlook, Mac Mail, Hotmail, Gmail, Yahoo! Mail, etc.)
- 6.3. How is email organized? (e.g. in self-created email folders, etc.)
- 6.4. How is email saved? (e.g. untouched in the email program, a copy in your PC, printed out in paper, etc.)
- 6.5. Are email and paper correspondence managed together or separately?
- 6.6. Do you use address books?
 - 6.6.1. Please list address books

7. Digital Files Storage / Backup

- 7.1. Do you have a backup routine for your files / emails?
- 7.2. What media are used for backup files? (e.g. optical disk, hard disk, file server, web based backup service such as SugarSync., cloud service etc.)
- 7.3. How recently have you changed computers? Do you transfer files in your old computer to your new computer
 - 7.3.1. If yes, what types of files are transferred?
 - 7.3.2. Did you encounter any problems in transferring the files?
 - 7.3.3. Please list problems encountered.
- 7.4. Do you keep your old computers?
 - 7.4.1. Please list details. [computer name(s), operating system(s), version(s)]

- 7.5. Have you ever experienced a serious hardware failure (e.g. hard-drive crash, loss of files, accidental deletion)?
 - 7.5.1. If yes, were the files in the affected computer recovered?
 - 7.5.2. If no, would like us to attempt to recover the files as part of your donation?
 - 7.5.2.1. If yes, which ones?
- 7.6. Are any digital files stored in unusual storage media? (e.g. punch cards, 8 inch. floppy diskettes, etc.)
 - 7.6.1. Please list media types.

8. Work Habits

- 8.1. Can you tell us about your work habits of using computers / mobile device? (e.g. work online, work offline, use mobile, etc.)
- 8.2. Do you share computer with other people?
 - 8.2.1. If yes, how are files created by different people separated?
- 8.3. Since a visual representation of working space may provide researchers additional information about your works, do you mind we take photos of your computer with surrounding space?

11. Privacy and security

- 11.1. Are some digital file types of a sensitive nature? (e.g. tax records, medical records, peer-review comments, letters of recommendation, student records, etc.)
 - 11.1.1 Please list categories of files and their restrictions.
- 11.2. Are there files that you would want destroyed?
 - 11.2.1 If yes, please provide details so that we can act upon when we encounter such files when processing your files.
- 11.3. Do any digital files require passwords?
- 11.4. Where are user names and passwords kept?
 - 11.4.1 What service / software are used to save them?
- 11.5. Do you use digital watermarks?
 - 11.5.1 Please list files and watermark rationale.

12. File Transfer Arrangement

- 12.1. Do you want to delete any files / re-organize the files before the transfer?
- 12.2. Are there files you would like to transfer to us later?
 - 12.2.1 Which files?
 - 12.2.2 When?
- 12.3. Can we take the original computer(s) or storage media?
 - 12.3.1 If yes, do you want the original media back once we've processed the donation? If no, we will destroy the original media once the donation is processed.
 - 12.3.3 What storage media will you deliver? [original hard drive, copy on external drive, dvd, cd, etc]
 - 12.3.4 If external drive, how is it formatted? (operating system, version)
- 12.5. Can we make a copy of the entire hard drives or just specific file folders?
- 12.6. Were these hard drives used for anything else, for example personal documents?

Digital Material Survey (Part II)

Note: This part of the survey is designed to be filled out by digital archivists regarding technical details of the tools used to create digital material.

1. Hardware

- 1.1 List the hardware configurations of each computers / mobile device. (e.g. manufacturer, model no, cpu, ram, hard drive capacity, video card, etc.)
- 1.2 Find out if the computers have USB ports or CD writers which could be used to copy the digital files.

2. Software

- 2.1 List the operating system and other system software with version number, installed in all the hardware.
- 2.2 Check if system date and time are set correctly. List the time zone used, if any.
- 2.3 With the help of the donor, list the main application software, with version no., used to create digital files.
- 2.4 If Microsoft Office is used, find out if the “User Name” field is set to the name of the donor. Find out similar setting for other main application software used.

3. Internet Access

- 3.1 Find out if the digital archivist can use the Internet access in the donor’s office using the digital archivist’s portable computer?

4. Networking

- 4.1 With the help of the donor, confirm if the computer is connected to file servers. Confirm if the donor save files in the file server. How much file server space is used by the donor?

5. Security

- 5.1 With the help of the donor, confirm if login is required to access desktop computers / mobile devices?
- 5.2 With the help of the donor, confirm if a digital certificate is used by the donor to login / sign digital files / encrypt digital files?
- 5.3 With the help of the donor, confirm if digital files are encrypted?

6. Comments (per section and per survey)

Building the Business Case for Digital Preservation

Neil Grindley

JISC

Abstract

This paper will describe work that JISC is funding to build an evidence-base of material that will provide a range of organisations with practical and plausible reasons for investing in digital preservation. The SPRUCE Project is leading on the work and is using hackathon-type events and other enabling measures such as disbursing small grants and providing practical technical advice, in order to better engage with stakeholders and elicit accounts of the real challenges that digital collection-holders face on a day-to-day basis. Complementary JISC-funded work in areas such as ‘sustainability’, ‘value’ and ‘costs’ will allow the SPRUCE project to build a nuanced and multi-faceted series of business cases. Case-studies arising from this project will be valuable in their own right, but the genuinely novel aspect of this work will be the assembly of methods used to arrive at convincing and usable arguments to support engagement with, and investment in, digital preservation.

Author

Neil is a Programme Manager at JISC with responsibility for digital preservation. JISC is an organization that funds and supports technology-related projects and services for the UK Higher and Further Education sector and is influential within and beyond the UK as an innovative agent of change. Neil is also currently a board member for: the Digital Preservation Coalition; the Open Planets Foundation; and the Alliance for Permanent Access. Previously, Neil has worked on projects supporting the use of advanced ICT methods for humanities research and was the IT Manager at the Courtauld Institute of Art.

1. Building the Business Case for Digital Preservation

1.1 The Problem

Based on a great many national and international activities over the last ten or fifteen years, there are reasons to suppose that digital preservation is now a well-described problem capable of drawing on a useful range of resources and tools to support practice. It is also apparent that a number of different international organisations and agencies have made a significant impact with helping practitioners across various sectors tackle the challenges inherent in the long-term management of digital information.

However, whilst much has been achieved, a gap and a challenge persist. Whilst the vast majority of organisations are now confident and willing to acknowledge—through a mixture of intuition and logic—that long-term access to (and therefore preservation of) digital materials is an area where thinking, strategy, and perhaps even policy is going to be increasingly vital, insufficient evidence currently exists to help them fully articulate what the nature, scale and scope of the threat to their digital assets is likely to be. To put it another way, they don’t yet have the means or experience to accurately gauge the *value* that digital assets represent to their organisation, and consequently don’t know to what extent digital preservation will serve the long-term financial and strategic interests of the organisation. The question thus becomes, ‘what is the business case for digital preservation?’

JISC has a remit to help UK universities and colleges ensure that they are managing their digital assets in the most effective way possible and part of this task is to support practitioners to be able to make

the case (either to internal decision-makers or external funders) for ongoing investment into digital preservation activities. JISC has therefore recently funded a collection of projects with the objective of building an evidence base that will support practitioners to make that case. The main purpose of this paper is to describe the assembly of methods by which the empirical evidence is being gathered and a useful by-product is to list some of the stated impacts that the projects have identified, which can in turn be assessed for the benefits that have accrued to the organisations involved.

1.2 Obstacles and Methodology

Any considered process to gather information in support of the business case for preservation meets several early obstacles. Firstly, it is clear that organisations will normally be reluctant to freely admit any failures to properly manage their digital assets. Reputation is an important currency for all organisations, public or private, but where the organisation has a remit to act on behalf of the public, and more especially where it has a trusted role around the stewardship of objects, reputation is a very important issue. So it will be difficult to identify and then subsequently disseminate stories where identifiable individuals in actual organisations have made inadequate attempts to manage their digital objects.

Secondly, it is difficult to gather information in the form of case studies and data from people with the simple proviso that the ‘community needs them’, or that it will ultimately be of benefit to the whole community that an evidence-base has been amassed. In the main, people need more motivation and more evidence of the near-term benefits to them to really engage with a process such as this.

Thirdly, and rather fundamentally, the fact that the digital preservation community is alert to the dangers of ignoring digital preservation does not mean that the broader community is necessarily convinced either by its requirement as an activity; or by the terminology and methodology proposed by those advocating digital preservation. This limits the potential sources of information from which the evidence base can be built.

To try and tackle these and other obstacles, JISC issued a call for proposals in September 2011¹ that invited ideas for 3 different types of project. The first strand called for a single project to ‘inspire, guide, support and enable UK HEI’s [and potentially other types of organisations] to address preservation gaps; and to use the knowledge gathered from that support work to articulate a compelling business case for digital preservation’. A mass of anecdotal evidence suggests that professionals working within organisations do not generally want to fill in surveys, or answer questions, or generally meditate *about* preservation. What they really want to do is get on and do something practical about their problems, ideally with recourse to accessible and expert support (on-site if possible) enabling them to tackle some of the specific problems that they are facing in their own institutions relating to the long-term management of information.

The objective, therefore, was to setup a support and enabling project that *actively* engaged and helped stakeholders with problems. This would in turn implicitly facilitate the exchange of information necessary to build the business case evidence base. In effect, the provision of case studies, use case information and stated requirements from stakeholders becomes the *quid pro quo* for the support and assistance given by the project to them. To further enhance the project’s chances of getting engagement from stakeholders, a proportion of the grant (the suggested amount was 20%, i.e., £50,000 from the total

¹ “JISC Grant Funding 12/11: Digital Infrastructure Portfolio,” last modified August 22, 2011, http://www.jisc.ac.uk/fundingopportunities/funding_calls/2011/07/grant12_11.aspx.

award offered of £250,000) was designated as an enabling fund from which the project could itself make small awards of up to £5,000 to institutions that wanted to pursue preservation objectives that might usefully serve as exemplars to the rest of the community.

The enabling grants added a fourth and potentially rich layer for the project to gather appropriate material to use in construction of the business cases. Going from the least to the most active engagement the project can gather data via the following routes (or as a by-product of the following activities):

1. Desk research and literature reviews
2. Interviews, discussions and questionnaires (this activity was not ruled out as a strategy)
3. Active support, advice and guidance to organisations with preservation challenges
4. Enabling grants to cover costs of preservation development activities

The second strand of activity in the JISC Call invited proposals for short projects to address a specified digital preservation problem and to do so within a period of 4 months (Nov 2011 – Feb 2012). The purpose was to set number of projects running that would provide the principle support and enabling project with a small existing stakeholder group which it could work with. It was also meant to ensure that early case study material emerged for inclusion into the evidence base, which would test assumptions around what sort of activity provided the most useful evidence, what kind of output and data would be required, and how effective the whole perceived mechanism would be for achieving the objectives of the call. A total of £75,000 was available to fund between 5-7 projects, attracting from between £10,000-£15,000 each.

In addition to helping the support project, these ‘*Active Case Studies*’, were an attempt to navigate around the first of the three obstacles defined above as hindering the gathering of evidence. Rather than just asking an organisation to tell everyone else about a digital preservation problem that they were facing (a passive case study), this measure also offered them a means to specify and implement a solution to their problem. Giving an organisation the means to fix and own the problem they have admitted to is a different proposition from simply asking them to announce that they have a problem.

The third strand of work was less focused on building the business case for preservation but was nonetheless an indirect attempt to address the third of the three obstacles. The purpose of this strand was to enhance the capability of organisations to more effectively include digital preservation as a component within the information management training and development courses and sessions that they organise for staff and students. This would in effect start to disseminate at least a basic level of preservation knowledge to more people within institutions and over time should result in higher levels of acceptance from more people about the principles, and perhaps the purpose (the business case even) of digital preservation. The call text makes the principle even more explicit,

Part of any embedding process for preservation should involve making appropriate aspects of it accessible and relevant to a range of people across the institution. By drawing out some of the more universally applicable issues around technology obsolescence, media refreshment, bit rot, long-term storage costs and digital forensics, it may be possible to positively influence the behaviour of staff and students within HEI’s to not only store their data more effectively, but to actively manage it, and therefore become more actively engaged with the value of that data and what it may or may not enable them to do.²

² “JISC Grant Funding 12/11: Digital Infrastructure Portfolio,” last modified August 22, 2011, http://www.jisc.ac.uk/fundingopportunities/funding_calls/2011/07/grant12_11.aspx (p. 27)

Given the level of funding (£100,000 for 3-4 projects of 9 months duration) this is clearly a limited contribution to what could potentially be a very large change programme, and one which would need a long timescale to measure demonstrable impact. It is however, an attempt to address part of the problem and it also again mentions the word ‘value’ and expresses the need to convey to stakeholders the idea that part of the challenge of purposefully engaging with preservation is understanding and acknowledging how much *value* digital materials have. This concept features in other areas of work that JISC has recently commissioned and some of this work will be referenced towards the end of the paper—particularly where it has direct relevance to building the business case. However, 12 months on from the invitation of proposals, it is now possible to examine and assess some of the activities that resulted from the call.

2. Projects

The enabling and support project was taken on by a consortium of organisations led by the University of Leeds and is called the SPRUCE Project (Sustainable Preservation Using Community Engagement).³ The principle idea behind SPRUCE continues on from an earlier JISC-funded project called AQUA (Automated Quality Assurance)⁴ which succeeded in defining and running events which brought together digital content specialists and technical developers to tackle real problems in real time (over the course of a 3 day event). These ‘people mashups’ allowed the content specialists to bring problems and questions to the event along with problematic or test data, and gave them a chance to sit down with a developer and work on solutions and ideas. Along with plenary discussion and presentation sessions and documentation that was captured as the event progressed on a dedicated wiki, these events received good feedback and more importantly, produced plentiful evidence that this was an effective way to accelerate the rate at which real problems encountered by those with responsibility for digital collections might be examined and tackled by technical digital preservation practitioners.

Over the course of just 2 events, the AQUA project produced descriptions of work on 24 different solutions to problems in seven categories.⁵ Admittedly these solutions were in different states of usability and functionality but it was clear that as an open and shared community activity and resource, there was great mileage in this form of cooperative problem-solving. Thus, the same approach was proposed for the SPRUCE project and the engagement required of stakeholders to build the business case was designed around further mashup-type events. At the time of writing SPRUCE has run one further event and has added further information about datasets, issues and solutions to the wiki page.⁶

What has become most obvious as a result of the AQUA and SPRUCE project outputs is that they are even more valuable as part of a larger community effort to align similar activities that are being instigated and managed by other entities such as the Open Planets Foundation (OPF),⁷ the SCAPE project,⁸ and the Digital Preservation Coalition.⁹ The information that has emerged from the JISC-funded

³ “SPRUCE Awards - funding opportunity for digital preservation,” last modified June 12, 2012, <http://wiki.opf-labs.org/display/SPR/SPRUCE+Awards+-+funding+opportunity+for+digital+preservation>.

⁴ “AQuA,” last modified July 16, 2012, <http://wiki.opf-labs.org/display/AQuA/Home>.

⁵ “AQuA Solutions,” last modified July 11, 2011, <http://wiki.opf-labs.org/display/AQuA/Solutions>.

⁶ “Digital Preservation Requirements and Solutions,” last modified May 8, 2012, <http://wiki.opf-labs.org/display/SPR/Digital+Preservation+Requirements+and+Solutions>.

⁷ “Open Planets Foundation: a community hub for digital preservation,” accessed September 5, 2012, <http://www.openplanetsfoundation.org/>.

⁸ “SCAPE: Scalable Preservation Environments,” accessed September 5, 2012, <http://www.scape-project.eu/>.

projects forms part of a larger body of material that has been brought together under the headings of ‘datasets’ (11 categories/60 items), ‘issues’ (19 categories/159 items), and ‘solutions’ (14 categories/80 items), all of which represents a rich treasure-trove of practical work that either practically supplies a working tool to fix a problem or issue, or at the very least flags up the work-in-progress thought processes that have been applied to a challenge, and may indicate directions for future work.

The URL is on a different section of the OPF wiki from the SPRUCE pages. See: www.wdl.org.

Whilst not yet specifically addressing the detailed objectives of the JISC call to collate and define business cases for digital preservation, it is clear that a great deal of information from a wide variety of sources is being marshalled to declare the type, frequency and scale of problems that organisations are facing in relation to the types of information management problems that digital preservation techniques are designed to tackle. The SPRUCE project is less than half way through its 2 year duration and is due to finish at the end of October 2013.

In the second strand of work, five Active Case Study projects were funded and constitute a diverse range of activities and tasks, all of which have different perspectives that will usefully build the evidence base to justify digital preservation.

2.1. The Carcanet Email Preservation Project

The Carcanet Email Preservation Project aimed to capture and preserve an important email archive and represented a good opportunity for the John Rylands Library at the University of Manchester to begin to engage with complex and emerging practice around a critical area of collections management.¹⁰ The final report is instructive in a number of ways and a good account of the various lessons that can be extracted from a short low-budget project. Sections 3.4 and 3.5 (p.18) are relevant for the evidence base.

2.1.1 Immediate Impact

- For the staff at Carcanet Press, significant progress has been made towards preserving their digital archive, and they are likely to have increased confidence in the Library’s ability to deal with further digital accessions in future.
- The project has ensured the ‘rescue’ for the archival record of a large body of research-rich material which was formerly at risk of loss; several of the files acquired resided only on the hard drives of Carcanet staff.
- The project marked the Library’s first acquisition of a substantial ‘born-digital’ archive, thus enhancing its ability to continue collecting modern archives in the digital age.
- It has improved the confidence of Library staff in their ability to acquire and preserve born-digital archive material, marking a vital first step in practical digital preservation.
- The project has contributed to the advancement of strategic goals, both for the Special Collections Division and for the Library-wide Digital Preservation Steering Group.

⁹ “Digital Preservation Coalition,” accessed September 5, 2012, <http://www.dpconline.org/>.

¹⁰ “Carcanet Email Preservation Project, Final report,” accessed September 5, 2012, <http://www.jisc.ac.uk/whatwedo/programmes/preservation/carcanet.aspx>.

2.1.2. Future Impact

- The project has resulted in a large digital archive which, although currently embargoed, will form a key resource for future researchers—not just those working in the field of literary studies, but also for historians, sociologists and others.
- The ability to deal with born-digital archives has the potential to enhance the UML's reputation as a collecting institution at an international level.
- The University's central IT Services have expressed interest in preserving institutional emails, and may draw on the experiences and outcomes of the project.
- The project will contribute to the growing dialogue about issues involved in the long-term preservation of email.

2.2 The Digital Directorate Project

The Digital Directorate Project, based at the Institute of Education, explored existing digital records management practice within the central administrative function of the organisation (the section producing most of the high-level governance and strategic planning records). The project created relevant documentation including a detailed retention schedule, an overview of relevant preservation metadata and procedures for the appraisal of electronic records. A final report is available.¹¹

2.2.1 Immediate Impact

The immediate internal impact of the project can be summarised as follows (Final Report, p. 8):

- It enabled the records management staff to establish new partnerships with key directorate staff (a highly influential group within the organisation).
- It clarified the organisation's internal committee structure which in turn improved information retention decisions.
- It opened a useful channel of communication between the records management staff and IT staff in relation to secure storage.
- It allowed project staff to become more familiar with the effective use of preservation metadata and the need to adapt standards to meet internal requirements.
- It afforded a chance for staff to develop procedures for the appraisal of electronic records, especially in relation to the acquisition of personal electronic archives.

2.2.2. Future Impact

The Future impact of the project is stated as follows (Final Report, p. 9):

¹¹ "The Digital Directorate: Digital preservation of governance records at the Institute of Education- Final Report," accessed September 5, 2012, <http://www.jisc.ac.uk/whatwedo/programmes/preservation/DigitalDirectorate.aspx>.

- A better understanding of internal committees will allow a substantial programme of de-accessioning to take place. This will be a large undertaking but will result in a streamlined historical archive and more space.
- A much larger programme will now be rolled out across other departments in the organisation based on templates and documentation developed in this short project.
- It will have a long-term and positive effect on the way the organisation uses metadata and how it tackles accession and cataloguing procedures.
- It will inform a forthcoming project in collaboration with the IT department looking at a permanent solution to the secure storage of semi-current electronic records.

It should also be noted that the Digital Directorate project was awarded an enabling grant from the SPRUCE project to continue aspects of its work.¹²

2.3. The Future Proofing Project

The Future Proofing Project was a collaboration between the University of London Computer Centre (ULCC) and the University of London (Archive).¹³ It worked with a simple toolkit of services and software and succeeded in plugging into a network drive to create preservation copies of core business documents that required permanent preservation. It purposefully set out to be a relatively simple intervention that made use of open source migration and validation tools. The case study sought to demonstrate the viability of the approach.

The report is a valuable and clear description of the work which should be of broad interest to a wide variety of people working in IT environments that are expedient rather than optimal. The impact of the project was as follows (Final Report, p. 34):

- The beginnings of a permanent record store
- The creation of a standard University of London position on digital preservation
- A holistic approach that works for many document/record types
- Costs savings, as storage is reduced
- Demonstrating the viability of an alternative approach to EDRM (electronic documents and records management)
- Good for disaster recovery and business continuity—puts set of core documents into a more portable safer place
- Shared drives not abolished, just managed better

¹² Details of all grants awarded to date are available at: <http://openplanetsfoundation.org/blogs/2012-06-08-spruce-makes-funding-awards-digital-preservation>).

¹³ “Future Proofing: enabling practical preservation of born-digital records - Final Report,” accessed September 5, 2012, <http://www.jisc.ac.uk/whatwedo/programmes/preservation/futureproofing.aspx>.

2.4. The Preservation of Publications in Education (POPE) Project

The Preservation of Publications in Education (POPE) Project¹⁴ aimed to reverse the effects of link rot on around 5,000 documents that made up the Digital Education Resource Archive (DERA).

The final report outlines the following future steps that will be taken as a result of the project work:

- A follow-on project has been specified to OCR image files representing text to render them machine-readable
- Work is being proposed to re-index the preserved publications using the London Education Thesaurus (LET) which would allow it to sit alongside content from other databases which had been classified under the same term
- A Preservation Strategy will be developed and plan for the IOE Library and Archives which will include ensuring that the material which has been preserved in POPE remains viable in the long-term
- The IOE Library has undertaken to ensure that all new publications which are accessioned to the library will also be preserved (where licences allow) in DERA. This is only possible because we have saved the time we spent fixing broken links and can now use it more profitably.

2.5. The Publishing Online to Preserve Scholarship (POPS) Project

The Publishing Online to Preserve Scholarship (POPS) Project¹⁵ grew out of a need for low-cost online journal publishing. Academic staff indicated that they wanted to be able to create online, Open-Access journals and sought advice from Learning and Information Services to see if a local solution was feasible.

Whilst digital preservation was not the principle objective of this project, it nonetheless represented a good opportunity to examine the aspects related to the sustainability of scholarly materials. The use of an ePrints repository platform brings material into an environment that has the potential to facilitate digital preservation and plug into relevant tools and methods. (See the JISC-funded Preserv project).¹⁶

Accounts of the third *Enhancing Capability* strand of work can be found in the four project blogs.

- DataSafe (University of Bristol) - <http://datasafe.blogs.ilrt.org/>
- SHARD (Institute of Historical Research & University of London Computer Centre) - <http://shard-jisc.blogspot.co.uk/>
- PrePARE (Cambridge University) - <http://preparecambridge.wordpress.com/>
- DICE (London School of Economics) - <http://lsedice.wordpress.com/>

3. Complementary Work

Building the business case for preservation is a complex proposition and there are no easily definable boundaries that can be placed around the topic. One of the critical factors that need to be addressed in any

¹⁴ "Preservation of Publications in Education (POPE) – Final Report," accessed September 5, 2012, <http://www.jisc.ac.uk/whatwedo/programmes/preservation/pope.aspx>.

¹⁵ "Publishing Online to Preserve Scholarship (POPS) – Final Report," accessed September 5, 2012, <http://www.jisc.ac.uk/whatwedo/programmes/preservation/pope.aspx>.

¹⁶ "Preserv2/EPrints Preservation Plugins 1," accessed September 5, 2012, <http://files.eprints.org/422/>.

consideration of business cases is the concept of sustainability and how to design strategies that lend assurance and permanence to an enterprise. One of the strong themes to emerge from work that JISC has collaboratively supported on the topic of sustainability—both in terms of the economic sustainability of digital preservation and access,¹⁷ and on the sustainability of projects and project outputs¹⁸—is that it will be difficult to design any long-term future for digital assets and services if their value proposition is uncertain or ignored. The concept of *value* is significantly linked to notions of ‘cost’ and ‘return on investment’ but also has overlaps and connections with other concepts such as benefits, impact, risk, efficiency, etc.

Another very important factor that is difficult to ignore when tackling this topic are the roles and responsibilities of those implicated in digital preservation, i.e., who needs to take ownership of the decisions that are required to manage and sustain digital assets over time, or to fund and maintain the solutions and services that are used for those management processes. An initiative that has tried to think through the relationship between some of these factors (mainly from the point of view of examining prospects for sustainability) is the Economic Sustainability Reference Model. This is a work-in-progress which is seeking to get broad community validation for a framework that will facilitate more thoughtful and sustainable design of measures to ensure long-term preservation and access.¹⁹

Further work on the reference model, on enhancing cost models for digital preservation, and on the other cost-determinants of digital preservation (e.g., risk, impact, benefits, etc.) is planned as part of a forthcoming European Commission funded project that is in the last stages of negotiation. This proposal should allow for intensive work to be carried out over a 24 month period to really help progress a raft of concepts and issues in this field and to join up the effort that a number of projects (including APARSEN,²⁰ TIMBUS,²¹ ENSURE²²) are already expending on different related parts of the conceptual jigsaw.

4. Conclusion

Building an effective evidence base on which to create convincing business cases for digital preservation will necessitate active collaboration with the community. This will best be achieved by a variety of different approaches spearheaded by an expert engagement project or programme that will lead and inspire people to contribute their stories and evidence. The enabling actions that occur in the process of collecting the evidence to illustrate the value of digital preservation may only be supported by modest financial awards of short duration but they appear to foster a desire within recipient organisations to implement lessons learnt more widely and to embrace more methodical policies and processes for information management across the organisation.

¹⁷ This relates to work carried out by “The Blue Ribbon Task Force for Sustainable Digital Preservation and Access,” accessed September 5, 2012, <http://brtf.sdsc.edu/>.

¹⁸ This relates to studies carried out by Ithaka S+R in association with the JISC-led Strategic Content Alliance, accessed September 5, 2012, <http://sca.jiscinvolve.org/wp/category/sustainability>.

¹⁹ An early account of activity is available from the “Unsustainable Ideas” blog, accessed September 5, 2012, <http://unsustainableideas.wordpress.com/economic-sustainability-ref-model-page/>.

²⁰ “APARSEN - Alliance for Permanent Access to the Records of Science of Europe Network,” accessed September 5, 2012, <http://www.alliancepermanentaccess.org/index.php/aparsen/>.

²¹ “TIMBUS - Timeless Business Processes and Services,” accessed September 5, 2012, <http://timbusproject.net/>.

²² “ENSURE - Enabling kNowledge Sustainability Usability and Recovery for Economic value,” accessed September 5, 2012, <http://ensure-fp7-plone.fe.up.pt/site>.

Requirements of a Remote Repository

Kevin Bradley

Abstract

In 2007 UNESCO Memory of the World published a report: "Towards an Open Source Repository and Preservation System: Recommendations on the Implementation of an Open Source Digital Archival and Preservation System and on Related Software Development." The paper argued for "digital simplicity" with the expectation that such a concept would lead to an uncomplicated digital repository, which could be implemented and maintained, by small group of talented individuals with a moderate level of technical expertise in a variety of situations. This paper re-examines the notion of digital simplicity and considers whether such an idea is achievable or whether it is nothing more than a form of technical nostalgia, seeking for an imagined simpler times, or the creation of a digital ghetto, with sub standard functionality. It also considers what remoteness means in the digital world, and what are the implications for the systems that exist in such physical, geographical and technical environments. This paper will draw out issues that arise in the seemingly contradictory situation where interaction with the complex is a necessity of participation, but simplicity becomes a requirement of sustainable digital preservation.

Author

Kevin Bradley has worked in the field of oral history, sound archiving and digital preservation for just over 25 years, primarily for the National Library of Australia (NLA). Currently Curator of Oral History and Folklore and Director of Sound Preservation, Kevin's previous appointments include Sustainability Advisor on the Australian Partnerships for Sustainable Repositories (APSR), a DEST-funded partnership, Manager of the Sound Preservation and Technical Services, and Acting Director of Preservation at the NLA. Kevin has been extensively involved in the preservation of general digital objects. For a number of years he managed Digital Preservation at the NLA and worked on the infrastructure project Australian Partnership for Sustainable Repositories. Kevin is a member of the UNESCO Memory of the World Programme, Sub-Committee on Technology.

1. Introduction

In 2007 a paper was published on behalf of UNESCO's Memory of the World Subcommittee on Technology. The paper was entitled *Towards an Open Source Repository and Preservation System*, and it surveyed the open source repository software area, at least as it was known to the author and the research team at the time, to determine whether a narrowly focused, fully OAIS functional low-cost repository system could be developed. The need for this is very clear; all over the world the preservation, digitization and documentation of social, cultural and technical knowledge is being undertaken in the digital domain. It is widely accepted that digital technology in isolation is a fragile media, and proper systems and technologies are required to maintain the content with integrity in both the short, medium and long-term. It is also equally obvious that the technology and systems to maintain that data in the medium and long-term exists primarily in more wealthy countries with large technical infrastructures. The paper stated that such an aim was possible.

This paper, like its predecessor, argues that we have a responsibility to develop appropriately scaled, responsibly safe, digital repository systems with preservation capability in the long-term, and data security in the short and medium terms as being one of its prime criteria. It also suggests that the technology and

information wealthy countries should develop open source, freely available, supported solutions for the issues described in order to enable nations, communities and others to exercise the right and ability to choose whether they wish to maintain and manage their own heritage materials in the digital domain.

2. The Urgency in the Need for a Repository.

All digital information, if it is to be managed and maintained, must be stored in an appropriate system, however there are a number of categories of digital materials with different urgencies associated with the need for a repository. As has been widely discussed, the prime motivation for the digitization of most two-dimensional physical media is access. And as most digitization in the current digital world is a form of access driven image digitization, most of the tools required for dealing with digital information needs are about organisation for access. Libraries and archives throughout the world have been creating digital images of their paper-based collections in order to make the material more freely and widely available, or at least to facilitate local access without damage to original materials. This process has a number of benefits, not least of which is a wide ranging and democratic access to information. The repositories and other digital tools which manage this are designed to simplify the processes and maximise the availability and discoverability of content.

Consequently, with regards to this particular scenario, the primary driver for building long-term data management capability into a repository is the economic benefit that might be enjoyed compared to the cost of read digitising an image collection. Even though it is also possible that in the long-term these image surrogates may well be the copies that out-survive their original sources, the risk of losing an image copy of a paper-based item and not being able to replace it in the medium term is quite low for most paper based material. And as long as paper-based records continue to survive in a robust form, digital repository managers are able to continue with their commonly risky data management policies in which the only likely loss is the cost of recreating the digital images, rather than the content itself. The risk based approach has effectively, if not consciously, been the way most small scale digitization systems operate. As a consequence, the open source digital repository environment has not developed this capability as thoroughly as they have access and dissemination.

However, most sound and audiovisual documents in the 20th century were created electronically, and these machine readable electronic records are now either obsolete or physically decaying due to the chemical failure of the carriers. The urgent need to copy and preserve the content of these items has been well documented and well described elsewhere.¹ In addition, most sound and audiovisual and still images created since the end of the 20th century have been created in digital form and stored on impermanent short term carriers. Most contemporary collections are acquiring original digital photographs, sound

¹ Dietrich Schüller, "Preserving the Facts for the Future: Principles and Practices for the Transfer of Analog Audio Documents into the Digital Domain," *Journal of the Audio Engineering Society* 49, no. 7/8 (July/August 2001): 618-621; Kevin Bradley, "Critical Choices, Critical Decisions: Sound Archiving and Changing Technology," in *Proceedings of the Pacific and Regional Archive for Digital Sources in Endangered Cultures*, 2004, ed. Linda Barwick et al. (Sydney: University of Sydney, 2003); IASA Technical Committee, *The safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy*, ver. 3, ed. Dietrich Schüller, 2005 (= Standards, Recommended Practices and Strategies, IASA-TC 03) (South Africa: International Association of Sound and Audiovisual Archives, 2006); IASA Technical Committee. *Guidelines on the Production and Preservation of Digital Audio Objects*, 2nd ed., ed. Kevin Bradley, 2009. (= Standards, Recommended Practices and Strategies, IASA-TC 04) (Australia: International Association of Sound and Audiovisual Archives, 2009).

recordings and video, without a clear pathway to manage them. So, where these documents are important, and it is necessary to maintain them, the repository which holds this material will be responsible for the only copies of these fragile virtual items. The future will not have access to most of the original carriers, but will have access to the content only through digitized facsimiles of managed data or files of the original object.

Large-scale data repositories and the National collections in the developed world manage the data in the same way as the banking environment that pioneered the approach; that is sophisticated and expensive storage management systems in an integrated data environment using multiple types of media and backed up by an IT team or contractors to undertake those tasks. However, these large-scale systems do not tend to scale down, and are not common in the smaller scale, remote collections, or collections in underdeveloped or developing countries because of the cost and the local infrastructure. So the urgent need for an open source repository system which incorporates preservation quality data management and backup is because those repositories to maintain these vital documentary materials need to maintain them as the sole source of otherwise lost materials.

Also equally clear is that the repositories of digital information are only as trusted, or as trustworthy, as the institutions that maintain and house them. So while it is important to create the technology that will allow long-term preservation, this will only be efficacious where the institutions that own and create those documents take on that role for the continued maintenance and existence of the content of those materials. Even though this paper argues for a widely available, reliable and low-cost system that could be easily deployed in a remote or technically less sophisticated environment, it also recognises the limitations of that aim. For a repository to be sustainable it must exist in associated technical infrastructure that is capable of supporting a functioning and sustainable system. Similarly there must be some level of technical knowledge and lease some recurrent resources, albeit at lower level, to make it sustainable.

3. The Outcomes of the 2007 Report

The report, *Towards an Open Source Repository and Preservation System*, was based on a set of industry understandings which by and large could be considered fairly common sense. Listed below are some of the points which informed the document. Underpinning this fairly broad statements was a plea for solving the digital preservation problem for small-scale repositories in a way that dealt with the majority of simple or basic digital objects. This statement was required because it was recognised that much of the work being done in digital preservation is solving the long-term sustainability issues for some very sophisticated technological objects. However the very complexity of the problems such expertise are trying to solve often excludes simpler systems and solutions which might be applied to the majority of formats that are used to document and record cultural materials generally. In the case of the digitization programme, the organisation has control over the formats and so even simpler solutions are potentially suitable.

- Create and store the content on a digital file in a format which does not apply any form of manipulation which causes data loss or loss of authenticity.
- Use a format which is widely implemented and supported, and preferably, though not necessarily, open or non-proprietary.
- Use a format that has a potentially long life (digitally speaking).
- Use a format that is most likely to have available migration pathways to the next format.

- Store enough metadata to be able to facilitate identification, access and preservation processes.
- Use a reliable storage format on at least two types of carrier.
- Make multiple copies, and check and verify them regularly.
- Plan to replace carriers and software as the market demands, and plan to migrate the content to the next type of reliable carrier.

The notion of digital simplicity, as discussed in the 2007 UNESCO paper, is appealing because it implies a simple long-term preservation plan. However, such a simplicity would be too great a cost if it were at the expense of providing access to significant content. Rather I wish to show that the notion of digital simplicity is embodied in a series of concepts regarding formats, content and their relationships, which when applied to long-term digital preservation increases the likelihood of survival of the content embedded in these formats. Digital simplicity can be considered at two levels; the level of the object which is being managed; the level of the repository which is being deployed to manage that object.

One of the main complexities, and one that has frustrated many thinkers about digital information, is identifying the actual characteristics of the digital object that need to be preserved. When looking at complex, computer-based or web deployed digital objects, it is difficult to distinguish between the information the object carries, the media or format that carries it, and the manner in which that object presents the information it carries. The media is, after all, the message, and as a consequence detaching the significant properties, or the essence, of such complicated digital objects is difficult. When the meaning embodied in an object is embedded in the objects, the relationships between other objects, and the way those objects are presented, the complexity multiplies. Virtually all modern web publication fall into this loosely defined category of complex materials. The technical implications of this is that being unable to define precisely what is the essence of an object, it becomes nearly impossible to design the systems which manage the long-term preservation of those characteristics.

The National Archives of Australia describes a number of categories of material which it takes into its digital repository.² They are listed below. Note however that they describe categories of material from the point of view of format or encoding information, rather than user categories.

- Archive files/wrapper/container
- Audio
- Computer aided design (CAD)
- Email
- Geospatial data
- Image
- Image - Vector
- Office documents
- Plain text
- Database tables, such as comma and tab-separated files (csv, tsv)
- Scripting files (such as Python, JavaScript, Perl, PHP)
- Structured Query Language (SQL)
- Video – including video stream, audio stream and container

² Allan Cunliffe, (2011). "Dissecting the Digital Preservation Software Platform, Version 1.0 (RKS: 2009/4026)," National Archives of Australia, accessed August 2012, http://www.naa.gov.au/Images/Digital-Preservation-Software-Platform-v1_tcm16-47139.pdf.

Of these audio, image and video might be construed as simple digital objects, and text may in certain circumstances be the bridge between simple and complex objects. There are a number of ways of looking at this; of which the first may be with regard to significant properties. JISC defines Significant Properties as follows:

Significant properties, also referred to as “significant characteristics” or “essence”, are essential attributes of a digital object which affect its appearance, behaviour, quality and usability. They can be grouped into categories such as content, context (metadata), appearance (e.g. layout, colour), behaviour (e.g. interaction, functionality) and structure (e.g. pagination, sections). Significant properties must be preserved over time for the digital object to remain accessible and meaningful.³

The significant properties of an audio file is in fact its audio characteristics, that is, if the process sets out to preserve an audio file it must preserve the complete technical/quality characteristics that embodies it. Similarly, this could be said of still images, and to some extent moving images as well. The significant properties are technically embodied in the quantitative measure of their quality. Text based documents are complex symbolic items that human society has been creating and adding to over the millennia. While it is quite possible to show examples that embody that complexity, our long familiarity with such objects allows us to make some pragmatic decisions about font, layout and spacing, and so read and preserve some fairly basic digital objects in text form. However, there is recognition that for most text objects, its significant property would be the information embodied in the text rather than other characteristics.

All other items in the list are a complex mixture of text, technology, relationships, files, standards and components that seems to defy a simple and automated way of being managed. So a simple digital object may well be one that the significant properties can easily, and technically, be separated from the media that carries it.

In addition to this, most of the formats associated with the simple objects described above have had a relatively long life, in digital terms. So, for example, image and audio formats used for preservation have been stable and standard for the past decade and a half. This is because the professional industry tends to resist change, unlike the consumer market which tends to embrace it. Likewise, professional involvement in the format means that there will almost definitely be a migration path from the old format of the new professional format; otherwise the market providers would find it difficult get the buy-in of the professional user.

These formats, all these categories of material, are often the primary source materials from which other more complex objects are created. They are also most often the primary documents for which most archives collections have the greatest responsibility, which carry the most significant information. So, a repository which exhibits digital simplicity would be one that deals with a limited number of files that exhibit characteristics described above.

Digital simplicity does not mean a technical ghetto, or a backwater which remains the same while the rest of the world changes, rather it implies an appropriate technology for the preservation of significant representational items which fall into a defined category, and which are the most significant for many archives.

³ JISC. “The significant properties of digital objects,” last modified 17 August 2010, <http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops>.

The Towards an Open Source Repository and Preservation System (2007) report made a recommendation for UNESCO's involvement "in open source software development on behalf of the countries, communities, and cultural institutions, who would benefit from a simple, yet sustainable, digital archival and preservation system," and to this end the Information for All Programme (IFAP), at the recommendation of the SubCommittee on Technology (SCoT), provided funding for the development of certain functionality, for which Archivematica won the project.

The Open OAIS defines Data Management, Ingest, Access, Administration, Preservation Planning and Archival Storage. The report also noted that most repository software was well developed in the areas of data management, ingest, access and administration, but little no work had been done in preservation planning and archival storage. The intention of the 2007 report was to create a repository which included the potential to manage All six of the OAIS categories.

To this end of the work undertaken by Archivematica produced a set of repository tools that that met that criteria in five of the six OAIS categories. Perhaps most significant was the inclusion of PREMIS metadata which enabled preservation planning in a way that was not possible with other open source systems. However, data storage remains an unresolved issue.

Most large-scale repositories buy large-scale data storage systems, and enter into some sort of arrangement with companies to supply the technology, software and data tapes. As has been previously stated, this sort of arrangement does not scale down well to a small-scale repository, represents a significant expense even on large, comparatively well resourced repositories and is much more expensive per gigabyte of storage when the repository data size is not great. The 2007 report looked to find a cost effective alternative.

At the time of writing (2007) the report stated that AMANDA, (Advanced Maryland Automatic Network Disk Archiver), an open source system which allows the IT administrator to set up a single master backup server to back up multiple hosts over network to tape drives/changers or disks or optical media, had potential to be incorporated for this purpose. Since that time the AMANDA system has been developed significantly and so has even more potential if it were incorporated into a repository system.

However, the data management landscape has also altered significantly, and it is worth reviewing at this time the options for managing data in a small-scale, tightly resourced archival digital repository.

There are a number of technical and socio-technical ways that a tightly resourced repository can ensure the medium to long-term integrity of the data it holds. We will consider some of those below; repository management partnerships, third-party providers, cloud storage, disk only and disk/tape systems. However, it is important to recognise that no media will last forever, and no data storage technology is reliable, rather it is the technical system that incorporates the technology that provides the reliability. The second aspect to recognise is that all data storage systems are at heart, a risk management technology, and the task is to find an appropriate technology which addresses appropriately the level of risk for the content the system manages and preserves and the environment in which it operates. All these technical systems are based in a particular environment, and we have to interact with that broadly defined system to fulfil all the tasks of a digital repository; collection building and management, preservation, and dissemination. In order to retain and sustain our collections we must create a continuous link between the present and the future, a link that ensures that the ongoing links in the chain of sustainability create a digital object that the future user can access in just the way we do now and does so in the specific social, political and economic circumstance in which it exists.

The archives, libraries and other knowledge keepers are dependent on these sophisticated technical systems and networks that are the pinnacle of our socio-technical capabilities. This means that the future

of our digital and digitized materials are linked directly to the ongoing and continuing support of the government and other organisations which fund our archives and collections. It is no overstatement to say that we are linked to a continuing civilisation.

More importantly, or perhaps just more urgently, we are linked to the ongoing economic and social stability of the societies in which we live and at risk when that stability is threatened. In the current, and seemingly ongoing international financial crisis, it is very concerning to realise that the greatest risk to our collections is not technical or chemical, is not related to magnetic domains, to optical surfaces or chemical solutions, but rather to the economic viability of the societies in which we live. Without ongoing funding, infrastructure and support, our collections are at risk. When considering the appropriate approach to collection management and preservation, we must take note of the social, cultural and economic environment in which that technical system operates. We must create systems that can be sustained in the economies in which they operate.

Repository management partnerships: it may well be appropriate to take all data held in a repository and back up that content in a partner repository that has the highly reliable, low risk, well resourced solution to the management of content. This effectively moves the risk out of the local storage environment and into the storage environment of another, remote place. The issues associated with this are by and large not technical, but rather those of the ownership of content and the property at cultural rights associated with it. Many national or nationalistic institutions tend to resist this, feeling quite rightly that they should be able to manage their own cultural information. Nonetheless, a very good relationship also has the effect of storing data in two quite separate areas, probably separate countries, which has an enormous advantage in protecting content against disasters. However, once crossing national borders and outside of national jurisdiction, other laws may apply to the content with regard to rights and ownership. Trust is probably the most significant aspect of this sort of relationship. However, when it comes to economic motivation, one of the flaws in the approach is that the repository that has the responsibility to maintain and manage the material is not the one that has the motivation to do so.

Third-party providers: There are an increasing number of commercial providers who will manage storage of digital content. Some of these provide very high quality, reputable systems, that replicate those found in the major national archives and libraries around the world, or indeed, exceed their capabilities. Some do not have those types of technical redundancies and capabilities. Besides assessing whether such suppliers meet the technical requirements of the repository, and whether the content they provide is managed with appropriate regard for the ownership of the original content and the legal constraints of the country from which they came, there is a need to negotiate a contract shall agreement regarding the quality of the data, an export and retrieval requirement in the case where either the storage company fails, or the content owner is unable to pay storage fees.

Technical solutions: “May all your problems be technical” Ancient I.T. worker blessing.

Cloud storage is the distribution of data and information across a number of storage environments, which are accessed, managed and controlled through a wide area network. The appearance of a single storage environment is created through a virtualiser, which keeps track of all the parts of the storage environment and renders it as a virtual storage system. Typically there is, or should be, a very high level of redundancy to manage the complexities of multiple storage environments. As can be seen, Cloud storage does not use any new storage technologies, but rather exploits the availability of a widely distributed high bandwidth network and uses the same storage technology as everyone else. Cloud storage hides the fixed

infrastructure and allows the marketplace to respond more flexibly to changes in demand for data storage. Cloud storage also can enable total cost savings because the fixed costs of maintaining a system are spread across multiple users.

However, the security risks associated with data stored in multiple environments, and discussed above for partnership storage, is multiplied by the number of storage environments, the crossing of national boundaries, and the transmission of that data across multiple environments. Like all aspects of data management these risks need to be considered and managed.

The same architectures and data models that underpins the OAIS apply to data stored in the Cloud in the management of long-term preservation of digital objects. There are also management risk and The National Archives of Australia have released a document entitled “A Checklist for Records Management and the Cloud”⁴ which highlights issues around security, authenticity, completeness, discoverability and access, and variability due to data management, copying and migration. David Rosenthal’s blog about the announcement of Glacier, and Amazon based storage environment, raises issues about latency and the cost of access over a given access threshold of around 5% per month (which means that retrieving an entire archive may take a long time, and/or cost a significant amount of money).

However, the most significant issue for an archive situated in an environment with limited infrastructure is the bandwidth of the connection to the network. The objects stored in digital repositories for long-term preservation tend to be very large, especially for audio and video items, and these items require high-speed large bandwidth networks to distribute the content around. As has been discussed widely, the connection speeds in developing countries can be very slow, and may be unreliable. In these circumstances a Cloud storage environment will be totally impractical.

Disk only storage: there is an increasing opinion that storing data on disk (spinning disk/hard metal) in RAID arrays is suitable for collections of significant and important data. Certainly hard disks manufactured in the past few years have better specification and performance characteristics than those from an earlier period. Nonetheless, numerous papers point to the failure of disk storage in critical environments, and the likelihood of data loss.⁵ Complex modelling of real-world failures point to a likely loss of data even when stored on disks configured in RAID. In big arrays, the likelihood of data loss is even greater as the multiple failure and replacement occurrences increased the likelihood of two failures in a single array. The conclusion of many such forums and discussions is to store your data in at least three different locations and to continue with the old advice of different types of media. It is also salutary to observe that most major archives still use tape and disk to maintain the integrity of the data.

⁴ “A Checklist for Records Management and the Cloud,” National Archives of Australia, 2011, last modified 2012, <http://www.naa.gov.au/records-management/publications/cloud-checklist.aspx>.

⁵ Robin Harris, “SSDs vs. Disks: Which Are More Reliable?” last modified January 2011, <http://www.datacenterknowledge.com/archives/2011/01/27/ssds-vs-disks-which-are-more-reliable/>; Robin Harris, “Google’s Disk Failure Experience,” last modified February 2007, <http://storagemojo.com/2007/02/19/googles-disk-failure-experience/>; Robin Harris, “Everything You Know About Disks Is Wrong,” last modified February 2007, <http://storagemojo.com/2007/02/20/everything-you-know-about-disks-is-wrong/>; Eduardo Pinheiro, Wolf-Dietrich Weber, and Luiz André Barroso, “Failure Trends in a Large Disk Drive Population,” in *Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST’07)*, February 2007 (Berkeley, CA: USEIX Association, 2007), 17-29, http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/archive/disk_failures.pdf.

Disk and tape, with multiple copies on tape, meets the requirements described in “Towards an Open Source Repository and Preservation System,” and remains a requirement for the small scale, tightly resourced repository described therein. As a consequence, there is a need to find funds to develop the incorporation of this into standard open source repository systems. The requirement is specific to those repositories in developing countries for all the reasons described above. However, it is not a requirement that is likely to be implemented for a well resourced repository even if it is using the same open source repository software. This is because well resourced repositories in the economically leading countries tend to be implemented and deployed within existing technical systems, using existing IT infrastructure and staff expertise. As a consequence, developing country is implementing open source repository software will not be able to piggyback off the developments required by their richer counterparts. As most repository deployment in these environments also includes new technical infrastructures, the cost barrier to establish a reliable sustainable repository is more significant in these cases.

4. Conclusion

There is still a need, especially in small scale, tightly funded digital repositories in archives in developing countries to find a solution to medium term data integrity, and to build that into the open source systems that are available. However, because the primary motivation for this comes from an underfunded sector, there will be a need to find grant funding to implement this necessary requirement as it will not come from the commercial sector as a response to its need, not the market generally.

Digital Forensics for the Preservation of Digital Heritage

Accountability for Archival Digital Curation in Preserving the Memory of the World

Wayne W. Liu

Department of Computer Science, Florida State University, USA

Abstract

Diverse parties empowered by the Internet and digital technologies are documenting events and publishing records, joining historians and archivists in collective memory formation. Democratized and inclusive collective memory provides robust checks and balances to official archives, and it reserves a way for future reprisals or reparations. Consequently, preservation takes a new meaning in this regard. In order to conform to historical mandates and eradicate falsehood, archival digital curation needs normative guidelines to include active selecting and updating in addition to passive safeguarding. We propose an accountability model to be the simpler and more succinct basis of new criteria and norms. We summarize four principles of accountability as: identification, authorization, attestation and retribution. Those principles help to transcend differences and resolve conflicts in historical cooperation to preserve more objective, true and just Memory of the World, which in turn will improve the accountability of nations and transnational organizations in World Politics.

Author

Wayne W. Liu received his Ph.D. in Computer Science in 2011, from the Florida State University. His dissertation, entitled “Trust Management and Accountability for Internet Security,” reflects his research interests in issues about Internet security and governance. He currently is an adjunct instructor at Thomas University. He loves God, dogs, arts, tennis and also has broad interests in law, literature, philosophy, and classical music.

1. Introduction

Memory of the World (MoW), like history or archives, is necessarily selective. Archivists take responsibilities to filter, determine and sort out a vast amount of information and disinformation to document events and reflect the minds, wills, deeds and background characteristics of those involved. Historians trust archivists to take custody of such records so they in turn can select, exploit, explicate and vitalize “historical facts” (Carr 1961). What happened in the past then can be preserved as wisdom or *collective memory* (Halbwachs 1992) to help mankind shape future. In theory, such educative purpose is noble and valuable as it helps humanity avoid reoccurring flaws and mistakes. But, in reality, undertaking such responsibilities can hardly be a straightforward endeavor that is totally objective or independent.

Historians may take pride in ethics and professionalism when standing up against outside influences to exercise best interpretive discretion on the authenticity, veracity or validity of written documents and verbal communications. Yet they may not be able to detect or overcome their own myopic ideologies or biased beliefs if any of such interferes with their work. Archivists share similar limitations in their duties and responsibilities. As they strive to draw right decisions and fair judgments for purpose-driven narratives, they may well be limited by the lack of funding or lack of access to restricted information sources. Important information may not survive the manipulation and control of powerful individuals, organizations or regimes if they have conflicting interests in dictating the sources and channels to get the outcomes they desire. In the past, limitations imposed by the powerful or self might have undermined

historians and archivists' efforts, causing history to be inevitably clouded or skewed, redacted or whitewashed to some legendary tales or reduced to propaganda of little or no trustworthiness, only to be rewritten by later historians or rulers, repeatedly.

Now, modern digital technologies and the Internet have given societies, particularly the archival and historical communities, the needed capability and an unprecedented opportunity to change that. Since, now, with the help of information technologies, the generation, preservation, protection and access of digital records can be made autonomous, ubiquitous and robust. Diverse parties now join historians and archivists in documenting events and publishing records. Their collective efforts to spread truths and shape memory across political, ethnical or other artificial boundaries cannot be easily dismissed by the powerful, giving historians and archivists a leverage to counter undue influences. Were historians and archivists in government institutions under political pressure, say, unable to maintain truthful accounts or stand by justice, such *collective memory* serves to provide checks and balances and reserves a way for future reprisals or reparations (Wallace and Stuchell 2011). So, knowing it or not, the nature of historians and archivists' tasks is no longer as vulnerable or subjective as before.

On the other hand, new technologies bring new challenges. The non-stop, unending flood of digital data from computers, networks, surveillance or mass communications, etc., are unlikely to be used again as a whole. Yet, it's prohibitively difficult to sift through such data streams and decide which pieces are, or will be, valid and important *records* (Duranti 2010) or which records have historical values can add to current archives (Cox 1994). So, instead of selecting and updating, new archival facilities are created to accommodate, not to use but to store, the huge amount of data that may vary in format, expanding in size. As digital technologies rapidly evolve, new expertise and resources are needed not only to deal with the new technologies but also to convert and transfer records stored with old technologies into the new formats. But, unless there is political incentive or financial support, ordinary archives may not get the preferential treatment to benefit from the state-of-the-art technologies but remain as is. Since it will be increasingly difficult for the archival institutions to keep multiple sets of expertise and resources in order to maintain these archives, these likely will risk obsolescence of technologies when the old technologies used to maintain these become unavailable over time (Hedstrom 2001). These archives are marginalized in a way as they are forced to retire from public access. Whereas those archives privileged to be converted because they are favored by the powerful likely will be showcased by the archival institutions, along with those new "archives".

When decisions on which archives to be transformed or transferred are based on political or technological dictations rather than on historical merits, and when archives are categorized and divided by media or formats rather than by contents, archives will be increasingly maintained by technicians oblivious of archival missions and values. Because the more resources devoted to maintain technologies, the fewer can be allocated to maintain contents, archivists and archival institutions are forced to retrieve from their active role in *collective memory* formation (Brown and Davis-Brown 1998; Wallace 2011) back to a passive shop-keeping role, yielding to budgetary manipulation or other undue influences of the powerful (Schwartz and Cook 2002) again. In the light of South African's apartheid-controlled archives (Sachs 2006), this may not be a healthy development for archivists, historians or for societies at large.

2. Archivists & Digital Curation

The trend of *digitization* in archival communities renews the awareness of *preservation* and gives rise to recent focuses on digital curation (Lee and Tibbo 2011; Yake, et al. 2011), which Elizabeth Yake (E. Yake, Digital Curation 2007) summarized as:

[T]he active involvement of information professionals in the management, including the preservation, of digital data for future use.

Based on Yake's brief description here, however, it is unclear how archivists or archival institutions are related to digital curation unless we can make a direct connection between archivists and *information professionals*. On exploration of such possibility, Lee and Tibbo have elaborated six dimensions in their *DigCCurr Matrix* of Digital Curation Knowledge and Skills, and suggest "archivists can take advantage of these connections to advance the archival enterprise" (Lee and Tibbo 2011). Specifically, they point out the connections are in *preservation*; i.e., archivists as information professionals must actively take responsibility of preservation in digital curation.

Nevertheless, the notion of active involvement in Yake's description seems to suggest archivists to participate more in the technological aspects of digital curation activities rather than focusing on their historical roles and duties. Most of the current discussions on digital curation and the fledgling education programs and academic curriculums also seem to presuppose a normal or political, social and ethnical stable environment to develop normative disciplines that focus more on technological or managerial knowledge and skills and less on the moral or ethical aspect of digital curation as if it's not an issue. Whereas the case of South African's apartheid-controlled archives (Sachs 2006) suggests that ordinary archival responsibilities conceived under such presupposition may not be sufficient to resist or counter corruptions in political, social or ethnical unstable environments where archivists and archival institutions are expected to play a significant role as a witness or monitor, or a historian, to stand up against the atrocity or injustice of the powerful, to witness and preserve evidence and proof for future recourses, reparations or retribution. We think talking about digital preservation in terms of how, without focusing on what to preserve and why, is a wrong way to answer the challenges of digital technologies. In fact, it probably can help archivists to really take the opportunity of technologies if we put focus on what and why when talking about how to use digital curation to make archival work more independent and meaningful.

Specifically, we think information professionals should serve archivists, not the other way around. Yes, archives can be part of digital curation; and an important part if you will, but an archivist's responsibility in its fullest and truest sense is much more than an information professional's responsibility. When archivists involve in digital curation, we should consider the meaning of preservation in terms of preserving the true and just Memory of the World as opposed to preserving everything. Archivists and archival institutions should not be dictated by insurmountable amount of digital data or the perpetual upgrades and updates of modern digital technologies. Rather, archival missions and values should reflect collective memory in general and the Memory of the World in particular. To this end, one important issue of digital curation---the controlled destruction of archives---remains not fully addressed. The *Reference Model for an Open Archival Information System (OAIS)* (Consultative Committee for Space Data Systems 2012), for example, in describing components and services required by archival institutions does not mention archive selection, appraisal, disposition, destruction or removal (Lee and Tibbo 2011).

As Pierre Nora has pointed out, using memory as a metaphor for archival work, the natural formation of memory is “remembering and forgetting” (Nora 1989). He even said “the indiscriminate production of archives is the clearest expression of the ‘terrorism’ that hijacks memory, and “while amateur archivists tend to keep everything, professional archivists have learned that the essence of their trade is to exercise controlled destruction” (Nora 1989). There is a difference between “keep everything” and “produce archives”. And the imperative of our epoch is the latter, not the former, or at least not just the former. Archives are not just storehouses or junk-dumps where you put some superfluous, no longer needed yet maybe can be recycled, objects there just in case these will be of some use again one day in the future. The true meaning of *preservation* is not to keep everything.

But, in practice, maybe it is! If keeping everything is the only way to preserve something. Here we see how digital curation can come to play and be used by archivists for historical merits. It is not important whether you keep everything or not. What’s important is there are “things” that must be preserved by archivists and archivists can use digital curation to achieve that preservation. Such preservation against, say, intentional erasure by the powerful must be done consciously. You almost can sense the danger when an archivist hides that something in a junkyard or storehouse, or in an archive, in order to safe-keep it as evidence or proof. This is what we need in archivists and archival institutions in case political or social corruptions make it extremely important to keep that evidence or proof. But those are extreme cases, i.e., we must see corruptions as exceptions not the norms. In normal situations, when archivists and archival institutions are not under undue influences, they should not store everything if that pollutes or muddles up archival presentation hence dilutes the values of archives and the effectiveness of preservation thus affects their more important historical mission, function and purpose, which is, in our opinion, to truthfully reflect the collective memory of social justice.

3. Archivists’ Responsibility

Archivists are unique among information professionals as they are hired specifically for preservation so they have the privilege to contribute to a “higher” cause--the preservation of collective memory. For this historical responsibility, competent archivists shall use digital curation with discretion to guide their daily responsibilities in promoting the cause. They are not confined to pragmatic doctrines, like keeping everything or not, because they need flexibility to use curation for such purposes. And this is true for all archival activities. Although archivists differ from other professionals in this regard, they do carry out duties and responsibilities; do conform to professional disciplines and ethics to be *responsible* in decisions and actions with loyalty and obedience, just like other professionals do. As for their exercising of discretion, this won’t be a problem since under normal circumstances their professional roles do allow some autonomy, delegation and maneuver. So, usually archivists should be able to work “wisely” as information professionals and follow professional disciplines and ethics to fulfilling their daily job requirements and the higher calling at the same time to contribute to the cause of preserving collective memory.

But exercising discretion can be a problem under abnormal circumstances, where *exceptions* occurred such as social or political injustice, atrocities, corruptions, or deceits, etc. When archivists’ historical responsibility calls for actions to serve the higher cause in such situations, those powerful individuals, organizations or regimes who committed the wrongs may block the calling in order to hinder the preservation of truth and the eventual justice. At this point, archivists are on their own, their professional disciplines and ethics won’t help because these likely were not developed to deal with such

situations. The professionalism may in fact conflict with the call for actions. So, archivists are left alone to make a choice.

They can either keep their disciplines and ethics, but give up their discretion or be oblivious of the wrongs as if nothing bad had happened but content to be *responsible* to work diligently, loyally and obediently, hence secure their jobs but fail their historical mission. Or, they can divert from the now inappropriate disciplines and ethics and realign their sense of duty and responsibility with the historical calling, to use discretion and their proficiency in digital curation to resist, counter, or record those exceptions, willing to be held *responsible* for disobedience or disloyalty in order to be *accountable* for their historical role but, then, may risk their jobs. Facing such a dilemma, nothing in the archivists' professionalism can guide their choice but their own subjective rationality, commonsense or conscience, which may not be reliable due to conflict of interest. So, to be *accountable*, or *responsible*? That is the ultimate test of archivists.

As nothing in archivists' professional repertoire gives them the needed justification or guidance to fulfill their historical calling during exceptional trying times, they may fail to act when their roles are particularly important to detect, deter, document the events of atrocities, corruptions or deceits, etc. Such professionalism obviously is inadequate to serve the mission of preserving collective memory as humanity's last line of defense for truth and justice. So, can we design archival curriculums to include the exercises of, say, disobedience and disloyalty, which are justifiable under the assumption of *exceptions*, not the norms, so that archivists' *responsibility* won't make them susceptible to political, social, ethnical or other undue interferences? We think the problems persist even we can do so. Because we will still rely on the archivists' subjective senses to tell if exceptions had occurred and there are undue influences trying to hinder the historical responsibility thus warrant disobedience and disloyalty. As such senses are subjective; archivists may have conflict of interest causing their subjective senses unreliable. Furthermore, those partake in documenting events and publishing records now include possibly untrained, unsophisticated and sometimes unscrupulous individuals or rogue organizations and regimes. They are the new "powerful" in the sense that now they have power to purposefully pollute or muddle up our collective memory rather than safeguarding it if they, say, do not and will not conform to professional or whatever disciplines. Although scholars have generally affirmed the values of unofficial documents and records (Josias 2011; Wallace 2011), there is yet a comprehensive and coherent strategy for archivists to integrate unofficial accounts into collective memory or to maintain their archives' integrity to reflect the collective memory. After all, who can disqualify others, or has authority and legitimacy to decide what shall be included or excluded in the collective memory?

What we need is a fundamental, simple, and very succinct set of principles as the basis of a framework upon which we can have norms and criteria applicable to both normal and exceptional situations to be used by professional and layman archivists alike to guide, hence can justify, their decisions and actions and allow them to reason within themselves and reason with each other; such that they can become reasonable and hence responsible and accountable. Such accountability and responsibility help them trust, communicate, exchange and understand ideas for archival work, to reach agreements to transcend cultural, social or religious differences but summon up their essence and use it to resolve political, ethnical or economic conflicts, to help each other improve each self, to work together to achieve historical cooperation in preserving true and just Memory of the World, which is the collective and accumulated wisdom of mankind that reflects the ultimate truth and the utmost justice. To this end, and since digital curation and digital forensics seem to face similar challenges (Diamond 1994; Duranti 2010; Lee and Tibbo 2011), we propose the same fundamental accountability principles we use to guide

systematic approaches in digital forensics engineering to be applied to sustain this historical-archival paradigm.

4. An Accountability Model

It is much debated among scholars about what *accountability* is exactly and how can we construct and enact accountability (W. W. Liu 2011). Since accountability is such a virtue that has been (mis)characterized as elusive or having a *chameleon* quality (Sinclair 1995), it is often misused to mean something else like blames, liability or punishment, etc. On the other hand, responsibility seems to be well understood and people seem to use the word all the time seemingly without causing any confusion. Nevertheless, as Gardner (Gardner 2003) has pointed out (without using the word), in its core sense responsibility is accountability; both are our *reasoning* capability as human beings. Thus, for us, *being responsible* in this sense is exactly the same as *being accountable*, i.e., we give reasons to respond or justify, or account for our decisions or actions; even in the former case this may be mandatory while in the latter case it's voluntary. But we don't usually use the words this way.

In our common usage of the words, there is a subtle difference between "holding someone responsible" and "holding someone accountable" that the former implies power and control while the latter implies democracy. For example, our superiors in an organization have power and control so they can hold us *responsible*, to lay specific blame, punishment or liability on us if we violated a rule or displeased them. On the other hand, we can only hold our leaders or superiors *accountable* for something they did not do right or didn't do, but they did not violate any law either. So they, not us, still have the power and control. Oftentimes, such subtle differences disappear as we use the words. For example, if our superiors did violate a law then the law, not us, can hold them responsible; but we may say we hold them responsible because we reported it. Or, if so many people blame a leader and hold him accountable then eventually such democratic force may remove him from power. Or, he may resign in order to show people his accountability, which is a required quality of democratic leaders.

From an individual's point of view, *responsibility* seems to be a virtue for subordinates. It is about being responsible to instrumental authorities (superiors), with loyalty, obedience, and efforts to take care of mandates and necessities with professionalism in duties and obligations; whereas *accountability* seems to be a virtue of leaders. It is about being accountable for missions and values, with self-motivation in self-control and self-improvement of leaderships that measure up to commonsense, social norms, altruism, or other higher civil, moral intrinsic orders. But every individual can be both a subordinate and a leader at the same time. Because, as human beings, even some of us are the instrumental authorities, we all are subject to intrinsic authorities like humanity and morality. On the other hand, even the least among us can aspire to have the leadership quality. Responsibility is purpose-driven, objective, more concrete and specific but can be too narrowly delineated, inadequate if become reactive or passive; while accountability is self-motivated, authoritative and encompassing, but can be subjective, vague or elusive sometimes. So, we really need both to complement each other in our historical-archival paradigm.

In our previous work (Liu, Aggarwal and Duan 2009; W. W. Liu 2010), we implemented four accountability principles, namely: *identification*, *authorization*, *attestation* and *retribution*, in a trust management framework for Internet servers to leverage organizations' civil roles to improve accountability in their trust relationships with users, peers and authorities. Those principles also are crucial for servers to bring deterrence and recourse to enforce responsibility so they can trust better, putting reliance on responsible users and peers while holding rogue users or peers responsible. But these

principles are also crucial for servers themselves to account for others and become trustworthy leaders or allies. Thus those principles facilitate accountability in both its *holding to account* and its *giving account* aspects. Now, as the professionalism and ethics manifested in archivists' responsibility have shown inadequacy in the case of South African's apartheid-controlled archives, we believe these accountability principles are useful for our historical-archival paradigm to help archivists make decisions and help them judge or justify their actions in the time of conflict. Specifically, the guidance and justification lie in the principle of authorization, which is based on autonomous *deference*, *reasoning* and *qualification* (W. W. Liu 2011), as shown in Table 1.

Table1: Archival missions, values, and processes defined by accountability.

	Missions	Values	Activities
Authorization	democracy	legitimacy	deference
		protection	reasoning
		access	qualification
Identification	no-deceit	worth	appraisal
		truth	authentication
		knowledge	classification
		reality	verification
Attestation	justice	information	communication
		education	presentation
		evidence, proof	certification
		utility	exhibition
		trust	reputation
Retribution	improvement	deterrence	punishment
		reparation	reconciliation
		correction	rehabilitation
		remedy	liability
		recourse	reprisal

Accountability in a democratic society is marked by deference to the higher, intrinsic, authority and legitimacy (Gardner 2003), unlike responsibility that is marked by deference to instrumental authorities such as rulers or enforcers. Thus, the way accountability works is different from the way responsibility works. The latter is like a command chain running from the top down and stopping at whoever cannot reject the duties or liabilities, whereas the former actually works in a reversed way. Oftentimes, accountability starts from someone *not* at the top who is willing to give account---e.g., a *whistleblower*---and it reaches upward, to get as many higher-ups as possible. So, if archivists use the authorization principle to answer the historical calling and take actions to document the wrongdoing of the powerful,

such decisions and actions are justifiable by the missions and values of the archival institution under the premise of accountability.

From Table 1, authorization primarily serves as a basis for judging archivists' decisions and actions under normal or exceptional circumstances, the other accountability principles are directly related to archival missions, values, and curatorial activities. Identification, for example, is directly related to archival selection and appraisal. Upon acquiring or receiving an object, archivists must actively determine its historical value and classify and categorize it before accepting it into archives. To have a place in the archives, any object not only must be recognized of its historical worth but also need be authenticated and verified before archivists attest to its identity. Once have a place in the archives, objects are maintained and protected in a way to allow safe access by the public. Public's opinions and feedbacks then may be collected and analysed to re-evaluate the objects. Attestation and retribution also apply directly to archival activities such as archival disposition, destruction and removal (Lee and Tibbo 2011).

On further thought, how about the missions and values of archival work? The four accountability principles have intrinsic values to reveal these. Table 1 shows how accountability principles define the missions and values of archival activities. Our assertion is that those accountability principles help to preserve true and just Memory of the World and that in turn will improve the accountability of nations and transnational organizations in World Politics. These four principles not only are effective means as guidance or justifications to help archivists but also can be an end in themselves. Put another way, maybe accountability is the essence of history and archives, preserved in the collective memory to serve the educative purposes of helping mankind shape a better future?

5. Conclusion

As the Internet has stimulated new ideas and new ways of communications to spread and share those ideas across traditional boundaries, it has decreased the efficiency of traditional power mechanisms used by old societies and organizations in accounting or law enforcing and regulatory supervision. On the other hand, the Internet and new technologies improve the awareness of constituencies on their potential inputs and influences on organizations' policies, actions and public relationships (Pruzan 1998); these bring about a wave of changes and new thinking on how to manage complex social systems. The vital, new perspectives emerged in business corporations and governments have shifted organizations' focuses from controls of fiscal and functional responsibilities to autonomy with ethical and managerial accountability. Archival institutions certainly can benefit from such changes too. The four principles of accountability not only are applicable to curatorial activities but also can help archival institutions' governance and public relations. The principles of accountability actually define the missions and values of archival work. We believe every archival institution ought to be an institution for, by, and of accountability. Archivists then can recognize, align with, and be accountable for the missions and values of archival endeavors to obey the intrinsic authority and legitimacy, to uphold social justice with their historical roles independent of temporal political, ethnical or other artificial powers, or instrumental authorities.

References

- Brown, R. H., and B. Davis-Brown. "The making of memory: the politics of archives, libraries and museums in the construction of national consciousness." *History of Human Sciences* 11, no. 4 (1998): 17-32.

- Carr, E. H. *What Is History?* New York: Random House, 1961.
- Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*. 650.0-M-2. 2012.
- Cox, R. J. "The Documentation Strategy and Archival Appraisal Principles: A Different Perspective." *Archivaria* 38 (1994): 11-36.
- Diamond, E. "The Archivist as Forensic Scientist—Seeing Ourselves in a Different Way," *Archivaria* 38 (1994): 139-154.
- Duranti, L., and B. Endicott-Popovsky. "Digital Records Forensics: A New Science and Academic Program for Forensic Readiness." *Journal of Digital Forensics, Security and Law* 5, no. 2 (2010): 45-62.
- Gardner, J. "The Mark of Responsibility," *Oxford Journal of Legal Studies* 23, no. 2 (2003): 157-171.
- Halbwachs, M. *On Collective Memory* (originally published in 1941, edited and translated by Lewis A. Coser). Chicago: University of Chicago Press, 1992.
- Hedstrom, M. "Exploring the Concept of Temporal Interoperability as a Framework for Digital Preservation," in *Third DELOS Network of Excellence Workshop on Interoperability and Mediation in Heterogeneous Digital Libraries*, 2001.
- Josias, A. "Toward an understanding of archives as a feature of collective memory." *Archival Science* 11, no. 1-2 (2011): 95-112.
- Lee, C. A., and H. R. Tibbo. "Where's the Archivist in Digital Curation? Exploring the Possibilities through a Matrix of Knowledge and Skills." *Archivaria* 72 (2011): 123-168.
- Liu, W. W. "Identifying and Addressing Rogue Servers in Countering Internet Email Misuse," in *IEEE/SADFE-2010: Proceedings of the 5th International Workshop on Systematic Approaches to Digital Forensic Engineering*, 13-24. Oakland, CA: IEEE Computer Society, 2010.
- Liu, W. W. "Trust Management and Accountability for Internet Security," Ph.D. diss., Florida State University, 2011.
- Liu, W., S. Aggarwal, and Z. Duan. "Incorporating Accountability into Internet Email," in *SAC'09: Proceedings of the 24th ACM Symposium on Applied Computing*. Honolulu, HI: ACM, 2009.
- Misztal, B. *Theories of Social Remembering*. Maidenhead, UK: Open University Press, 2003.
- Nora, P. "Between Memory and History: Les Lieux de Memoire." *Representations* 26, sp.issue (1989): 7-24.
- Pruzan, P. "From Control to Values-Based Management and Accountability." *Journal of Business Ethics* 17, no. 13 (1998): 1379-1394.
- Sachs, A. "Archives, truth and reconciliation." *Archivaria* 62 (2006): 1-14.
- Schwartz, J. M., and T. Cook. "Archives, records, and power: the making of modern memory." *Archival Science* 2, no. 1 (2002): 1-19.
- Sinclair, A. "The Chameleon of Accountability: Forms and Discourses." *Accounting, Organizations and Society* 20, no. 2/3 (1995): 219-237.
- Wallace, D. A. "Introduction: Memory ethics—or the presence of the past in the present." *Archival Science* 11 (2011): 1-12.
- Wallace, D. A., and L. Stuchell. "Understanding the 9/11 Commission Archive: Control, Access, and the Politics of Manipulation." *Archival Science* 11, no. 1-2 (2011): 125-169.

Yakel, E. "Digital Curation." *OCLC Systems & Services* 23, no. 4 (2007): 335-340.

Yakel, E., P. Conway, M. Hedstrom, and D. Wallace. "Digital Curation for Digital Natives." *Journal of Education for Library and Information Science* 55, no. 1 (2011): 23-31.

Automated Redaction of Private and Personal Data in Collections

Toward Responsible Stewardship of Digital Heritage

Christopher A. Lee¹ and Kam Woods²

School of Information and Library Science, University of North Carolina at Chapel Hill, USA

¹calee@lis.unc.edu, ²kamwoods@email.unc.edu

Abstract

In order to support digital heritage, collecting institutions must serve as trustworthy and responsible stewards of digital information. This requires not only selecting, acquiring and retaining valuable collections, but also providing appropriate access to their contents. Access provision involves data mediation to provide useful access points, to convey contextual information, and to ensure that private information is protected. Identification and redacting of private information is a significant challenge for collecting institutions providing access to born-digital collections. We describe work in the BitCurator project to provide collecting institutions with reliable, automated software and reporting procedures to address the above issues.

Authors

Christopher (Cal) Lee is Associate Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. His primary area of research is the long-term curation of digital collections. He is particularly interested in the professionalization of this work and the diffusion of existing tools and methods into professional practice. Lee edited and provided several chapters to *I, Digital: Personal Collections in the Digital Era*. He is Principal Investigator of the BitCurator project, which is developing and disseminating open-source digital forensics tools for use by archivists and librarians.

Kam Woods is Postdoctoral Research Associate at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He is currently the Technical Lead on the BitCurator project. Woods received his Ph.D. in Computer Science from Indiana University, Bloomington and his Bachelors with a special major in Computer Science from Swarthmore College. His research focuses on long-term preservation of born-digital materials; he is interested in interdisciplinary approaches that combine expertise in the areas of archiving, computer science, and digital forensics in enabling and maintaining access to digital objects that are at risk due to obsolescence.

1. Introduction

In order to support digital heritage, collecting institutions must serve as trustworthy and competent stewards of digital information. Responsible stewardship requires not only selecting, acquiring and retaining valuable collections, but also providing appropriate access to contents of collections. Access provision involves various forms of mediation, in order to provide useful access points, convey contextual information, and ensure that sensitive information is not inappropriately disclosed.

Modern computing devices often contain a significant amount of private and sensitive information. The information may appear embedded in document content and document metadata, within system files obscured by operating system arcana, and in traces of content held within local storage systems after interaction with services over a network. It may even inadvertently be retained on disk or in static memory after a device has been retired or formatted without being overwritten.

If collecting institutions do not improve their processes for discovery, identification and redaction of sensitive information, there are three major risks to digital heritage. First, collecting institutions could fail to be perceived as trusted actors who can responsibly care for digital collections. As a result, producers of digital content may be unwilling to transfer their materials to institutions for long-term care. Second, if the costs of processing collections are prohibitively high, then institutions are likely to acquire fewer collections than they would otherwise. Finally, reliance on labor-intensive manual procedures will ultimately result in rapidly growing backlogs of unprocessed material that are stored but not available for use. Lack of use is one of the most serious threats to long-term preservation of digital collections. Many of the tasks associated with identification and redaction of private and sensitive information can be simplified and performed with improved coverage and accuracy through the use of open source digital forensics tools that have been designed to work efficiently with large data streams, scale effectively on multiprocessor and multi-core hardware, support a common metadata schema for information transfer, and emphasize extensibility through the use of developer-friendly plug-in architectures.

2. Private Data in Digital Collections

For the purposes of automated analysis, triage, and redaction, we present the following definition of private data: any data that are personally identifying, could be used to establish the identity of the producer, establish the identity or personal details of individuals known to the producer (e.g., friends, family, and clients) or are associated with a private record (e.g., medical, employment, and education). This definition is specific to features one can extract from digital materials that link them directly to the donor or content producer. A donor could consider to be “private” a collection of documents or media which contain no personally-identifying information, but which he/she would not wish to disseminate publicly. Such collections of digital objects do not fit our working definition of private, but they may be considered *personal* or *personally identifying*.

We can further posit instances of features within born-digital objects that fall within our stated criteria, but do not require protection or redaction. However, when working with large (multi-terabyte or larger) collections of heterogeneous objects, it can be desirable to flag all potential instances of such data and perform filtering and triage after the fact.

2.1 Distinguishing between Private and Non-Private Data

Not all private data requires redaction, and the majority (in terms of bytes) of raw data held on media procured from private donors is likely to be non-private. Donor media obtained as disks or disk images of a modern bootable operating system, for example, will likely contain tens or hundreds of gigabytes of operating system and program executables, shared libraries, and documentation. Furthermore, common locations of private content—user-specific directories, storage volumes, and removable media—are just as likely to contain large collections of non-private content (e.g., PDFs downloaded from the web, commercial music and video content, and installation files).

Private data, under the definition provided above, can be associated directly with an individual. This can include, but is not limited to, social security numbers, credit card numbers, financial records, medical records, employment information, education records, passwords and cryptographic keys, and local and online account records. Private data may also be embedded in data carriers that are not inherently private, such as personal emails and other forms of electronic correspondence.

Collecting institutions deal with large volumes of information that do not meet this strict definition of private, but still requires identification, management, and protection. *Personal* and *personally identifying information* is not inherently private, and includes information that belongs to a specific individual (or group of individuals). Users are typically aware that contact names, telephone numbers, emails, and email addresses are personal and contain personally identifying (and in some cases private) information. Users may be unaware of other types of personally identifying information; for example, EXIF metadata within jpeg image containing geolocation data (GPS tracking information), and logs of user activity stored by an operating system over time.

The operating system on modern computing platforms will often store private and personal data in storage locations alongside non-private data. Depending on configuration and use, a wide variety of structures stored within the filesystem may include data that may be considered personal, personally identifying, and/or private. Logs created by installed applications can include usernames, full legal names, and other personally-identifying data such as addresses, birthdates, and passwords. Usernames and passwords and user account information may be stored in plaintext or using cryptographic techniques that have known attack vectors and readily-available exploits. Reconstruction of network traffic cached by certain applications can yield further vulnerabilities. Log information stored both by the operating system (for example, records of types of external storage devices along with serial numbers and timestamps), by applications (chat and social network logs and cached data), and even as a result of efforts by the user to redact or clear events on the system may also be used to reconstruct traces of personal and private data. Form data and cookies stored by web browsers are a well-known vector for such information. Less well understood (even, in many cases, by technically proficient users) is the fact that personal and private data can be recorded ‘inadvertently’ in hibernation and restore files used by modern operating systems to ensure stateful recovery from suspend events (such as putting hardware to sleep) and system crashes.

Not all traces of interactive use and other activities are private. We would contend that the following will generally contain non-private data:

- Filesystem type and organization (including partitioning info and existence of cryptographically-protected volumes)
- File modification times
- File names (except when those names contain personally identifying information)
- Names and versions of software used to produce digital documents
- Metadata about the hardware, operating system version, and security updates

Accurate identification of non-private data can be a critical aspect of preparing born-digital materials for long-term archival management and access. Specifically, it is often more efficient to cull non-private and non-unique data using known file hash sets (such as those provided in the National Software Reference Library¹) prior to performing deeper analysis on file contents. Identification of this type of data can also assist in building profiles describing and recording the environment in which documents were produced over the lifetime of the filesystem.

2.2 Identifying and Managing Private and Personally Identifying Data

Finding and categorizing private and personally identifying data in raw, heterogeneous data sources can be an arduous, labor-intensive, and error-prone task. As an example, string and regular expression

¹ <http://www.nsrll.nist.gov/new.html>

searches² are limited by prior knowledge of what is being sought, fail in the presence of extended character sets, and are often implemented in programs that do not attempt to parse known binary formats, compressed files or filesystems, or file metadata. Procedures that parse only the currently allocated areas of a filesystem may fail to find private and personal data that remains in ‘deleted’ (de-allocated but not overwritten) files, damaged areas of the filesystem, and ‘slack space’—data existing in blocks associated with allocated files that have not been completely overwritten (Garfinkel, 2010). Many tools that provide interactive low-level access to raw devices or disk images depend on extensive post-processing to generate useful reports on what is being observed, or depend on significant expertise and training on the part of the user.

Addressing the requirements of accurately identifying specific features in diverse filesystems and data formats can result in a patchwork of procedures and software mechanisms accumulated over years of effort and experimentation. This patchwork must then itself be curated, along with its associated training documentation and administrative overhead. This is a fundamental barrier to sustainability; knowledge of the system can become increasingly fragmented as training costs and complexity increase. These approaches also scale poorly with increased data volume, as each tool must typically be run independently over the objects in question.

These issues are not unique to collecting institutions, and have many parallels to issues encountered by digital forensics investigators. Garfinkel (Lessons Learned 2012) summarizes findings by Hibshi et al. (2011) that describe fundamental issues that law enforcement professionals have when dealing with raw, heterogeneous data collections:

They are general deadline-driven and overworked. Examiners that have substantial knowledge in one area ... will routinely encounter problems requiring knowledge of other areas ... for which they have no training. Certifications and masters’ degrees are helpful, but cannot fundamentally address the diversity problem as any examiner might reasonably be expected to analyze any information that their organization might possibly encounter on digital media (S82, emphasis added).

One of the underlying problems—that of working with heterogeneous forms of data from many sources—can be mitigated through the use of high performance, multipurpose software tools that consolidate functionality. Ideally, these tools should provide mechanisms for reporting on data in ways that are neutral with respect to a use case or workflow, allowing institutions to customize the technology for unique aspects of their respective institutional environment.

3. Applying Digital Forensics to Digital Collections

More than a decade ago, a report by Seamus Ross and Ann Gow (1999) discussed the potential relevance of advances in data recovery and digital forensics to collecting institutions. More recently, there has been an active stream of literature related to the use of forensic tools and methods for acquiring and managing digital collections. This has included activities at the British Library (John 2008), National Library of Australia (Elford et al. 2008), and Indiana University (Woods and Brown 2008; Woods and Brown 2009).

²String searches look for specific runs of characters within the data. Regular expressions are a more complicated way to identify particular patterns, allowing one to find, for example, given combinations of characters even if they appear in different sequences or are separated by other characters.

The PERPOS project at Georgia Tech has also applied data capture and extraction to US presidential materials (Underwood and Laib 2007; Underwood et al. 2009). A project called “Computer Forensics and Born-Digital Content in Cultural Heritage Collections” hosted a symposium and generated a report (Kirschenbaum, Ovenden and Redwine 2010), which provided significant contributions to this discussion. The Born Digital Collections: An Inter-Institutional Model for Stewardship (AIMS) project developed a framework for the stewardship of born-digital materials that includes the incorporation of digital forensics methods (AIMS Working Group 2012). The Digital Records Forensics project has also articulated a variety of connections between the concepts of digital forensics and archival science (Duranti 2009; Duranti and Endicott-Popovsky 2010; Xie 2011).

3.1 Acquiring Content from Born-Digital Media

When information is obtained from an individual or organization on fixed or removable digital media (such as hard disks, CD-ROMs, and USB drives), the collecting institution often does not know the state of the—whether they are physically damaged, whether the filesystems were left in consistent states by the producer or program which last interacted with them, or whether there are traces of previous (unrelated to the acquisition) filesystem activity remaining on the device.

Simply mounting a device on an appropriately equipped workstation and examining the contents of the filesystem in a desktop environment can pose significant risks to the consistency and authenticity of writeable filesystems (even when the device itself is mounted read-only). Likewise, many modern operating systems will ignore multiple disk volumes (mounting only the first volume), or fail to mount filesystems that are common on other modern platforms (or were common on a legacy platform).

In order to ensure that data from an acquired device remain unchanged (but can still be analysed thoroughly in their original state), the media can be *forensically imaged*. A bit-identical copy of the data on the media (known as a disk image) can be extracted using *write-blocking hardware*, which is connected as a physical proxy between the media-reading device and the host system and prevents the host system from changing any data on the original media. Typically, this disk image will be packaged with supporting metadata that describes the time, date, and other relevant information about the imaging process, using forensic imaging software that can write to an open imaging format such as the Advanced Forensic Format (AFF) (Garfinkel 2006, 2009), or well-documented commercial formats such as Guidance Software’s EnCase Forensic Image. Examples of such software include the open source tool Guymager and the commercial tool Access Data’s FTK Imager.

Bit-identical images of source media are useful for many reasons other than identifying and redacting private data. A donor may have forgotten where certain materials are on disk, whether they are password-protected or encrypted, or may have accidentally deleted important records. Access to the complete disk image provides a greater chance of recovery in such situations. Programs and documents located on the original disk may depend on fixed paths. The original filesystem may have damage that can be identified and repaired with access to logs and other system recovery data. Once a disk image has been created, one can automatically extract both general information about the filesystem, and instances of features corresponding to the types of private data outlined earlier.

Digital forensics workflows usually incorporate disk image packing formats, which include not only the raw data from a disk (including all original filesystem metadata from the disk) but also rich metadata about the capture process. Consider the following selected sample of metadata encapsulated in an AFF disk image extracted from a circa-1999 4.2GB external USB hard disk:

Table 1. Metadata Associated with a Disk Image as Part of an AFF Package

Segment	Data Length	Data
description	48	Quantum Fireball 4.2GB External
notes	16	000ECC21000831B9
badsectors	8	0 (64-bit value)
md5	16	C553 4AEA 0440 C75E 1B0A 508D...
sha256	32	954B A2B8 30F9 19C5 81F9 F546...
acquisition_seconds	0	= 00:09:46 (hh:mm:ss)

The complete AFF metadata provides a significantly more nuanced view of the capture process, but the information in Table 1 alone provides valuable reference points for future analysis; capture time, multiple cryptographic checksums for the raw image, the serial number of the original hardware (when available), and the number of bad sectors encountered.³

Both the filesystem metadata within the disk image and the supplementary metadata within the disk image package can be used to document provenance⁴ and chain of custody.⁵ Forensic formats typically “chunk” data, associating cryptographic checksums with each chunk. When compression is used, it is generally at the chunk level rather than at the level of the file; because of this, if partial data loss occurs (due to physical or logical failure on the storage medium), the undamaged parts of the disk image can be recovered with relative ease.

Note that analysis of disk images does not depend on the use of a custom forensic format; the majority of modern forensics tools will operate just as effectively on raw bitstreams (and even raw devices, should the need arise). In the following section, we discuss software tools and approaches being integrated into BitCurator.

3.2 BitCurator

The BitCurator project is a joint effort—led by the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS) and Maryland Institute for Technology in the Humanities (MITH), and involving contributors from several other institutions—to develop a system for librarians and archivists that incorporates the functionality of many digital forensics tools (Lee et al. 2012).

Digital forensics offers methods that can advance the archival goals of maintaining authenticity, describing born-digital records and providing responsible access (Woods and Lee 2012). However, most

³ Note that “Data Length” refers to the structure of the output item, not the relevant data value – in this case, no bad sectors were found.

⁴ Provenance “consists of the social and technical processes of the records’ inscription, transmission, contextualization, and interpretation which account for its existence, characteristics, and continuing history” (Nesmith 1999, 146).

⁵ Chain of custody is the “succession of offices or persons who have held materials from the moment they were created” (Pearce-Moses, 2005, 67). It can be ensured through control, documentation, and accounting for the properties of a digital object and changes of state (e.g., movement from one storage environment to another, transformation from one file format to another) throughout its existence—from the point of creation to each instance of use and (when appropriate) destruction.

digital forensics tools were not designed with archival objectives in mind. The BitCurator project is attempting to bridge this gap through engagement with digital forensics, library and archives professionals, as well as dissemination of tools and documentation that are appropriate to the needs of memory institutions. Much BitCurator activity is translation and adaptation work, based on the belief that professionals in collecting institutions will benefit from tools that are presented in ways that use familiar language and platforms.

BitCurator—and the efforts of many of the project partners—also aim to address two fundamental needs of collecting institutions that are not priorities for digital forensics industry software developers:

1. Incorporation into the workflows of archives and libraries, e.g., supporting metadata conventions, connections to existing content management system (CMS) environments. This includes exporting forensic data in ways that can then be imported into descriptive systems, as well as modifying forensics triage techniques to better meet the needs of collecting institutions.
2. Provision of public access to the data. The typical digital forensics scenario is a criminal investigation in which the public never gets access to the evidence that was seized. By contrast, collecting institutions that are creating disk images face issues of how to provide access to the data. This includes not only access interface issues, but also how to redact or restrict access to components of the image, based on confidentiality, intellectual property or other sensitivities.

Two groups of external partners are contributing to BitCurator: a Professional Expert Panel (PEP) of individuals who are at various stages of implementing digital forensics tools and methods in their collecting institution contexts, and a Development Advisory Group (DAG) of individuals who have significant experience with development of software. The core project team met with the PEP in December of 2011 and the DAG in January of 2012 to discuss the design assumptions and goals of the project. We have also received comments and suggestions from individuals in a variety of organizational settings. These various forms of input have helped us to refine the project's requirements and clarify the goals and expectations of working professionals.

The project is packaging, adapting and disseminating a variety of open-source applications. BitCurator is able to benefit from numerous existing open-source tools.⁶ The goal is to provide a set of tools that can be used together to perform curatorial tasks but can also be used in combination with many other existing and emerging applications.

In the following sections we briefly discuss risk-reducing practices for acquisition of content from born-digital media, and examine in detail how modern digital forensics tools can be used to ensure that private and personally-identifying data is quickly and accurately identified and reported.

4. Applying Digital Forensics to Identify and Redact Private and Personally Identifying Data in Born-Digital Collections

A primary goal of BitCurator is to enable professionals at collecting institutions to rapidly and accurately identify private and personally identifying data. We facilitate this by using a toolset that is intended to be

⁶In addition to those discussed in the following text, BitCurator is also incorporating Guymager, a program for capturing disk images, and Nautilus scripts to automate the actions of command-line forensics utilities through the Ubuntu desktop browser.

simple to operate, constructs well-formatted human- and machine-readable reports on those data, and interoperates effectively with existing archival data management systems.

The operational aspects of research efforts conducted as part of BitCurator depend on a set of mature open source software technologies originally developed for digital forensics and law enforcement investigations. The technologies we depend on include bulk extractor and fiwalk (developed by Simson Garfinkel), The Sleuth Kit (developed by Brian Carrier and Basis Technology), and sdhash (developed by Vassil Roussev). These projects, their functions, and our adaptations are described in the following sections.

4.1 Filesystem Analysis

While it can be useful to mount and explore a filesystem interactively on a host machine, it is often undesirable to rely solely on information gained this way. First, this is a risky activity if a hardware write-blocker is not used. Second, this will not yield information on deleted contents, unallocated space, and (often, without specialized software) secondary or alternate filesystems contained on the device. Finally, this form of investigation is time-consuming and prone to human error.

To report on the contents of the filesystem, we use fiwalk, a program originally developed by Simson Garfinkel and now integrated into The Sleuth Kit. Fiwalk processes disk images using the filesystem processing libraries in The Sleuth Kit, and generates results either in Digital Forensics XML (DFXML)⁷ or as human-readable plaintext. A typical run of fiwalk will produce some technical metadata on the operation of the program and the host environment, along with a complete walk of the file hierarchy—including volume information, directories, regular files (including lengths and block offsets within the disk images), and files which are no longer allocated (deleted).⁸ Additional uses and examples can be found at <http://afflib.org/software/fiwalk>.

As an example, consider the following sample of XML output of fiwalk being run against the legacy 4.2GB external USB drive referenced in Section 3.1 (a disk that includes real-world data and is part of the BitCurator Disk Image Test Corpus discussed in Section 4.6). This sample includes a single “fileobject” entry for a file that is orphaned—that is, a file that originally supported the operation of an installed program and is no longer used or referenced by that program. Such files are typically not visible to users interacting with a filesystem (Figure 1).

Note that this file is almost 50KB in size. In section 4.3, we provide a further example of how information extracted by forensic data analytics tools can be used to link private and personally identifying information to specific data objects (and specific byte offsets within those objects), particularly when they are “hidden” in this manner.

Using an existing Python module and supporting code distributed with fiwalk, the BitCurator project is preparing programs that can be run against collections of disk images to generate human-readable reports on information useful for triage and data sanitization tasks: for example, the number and type of filesystems recognized, and distribution and location of file types.

⁷ For a discussion of DFXML, see our section on Data Reporting.

⁸ Fiwalk currently recognizes those disk formats which are supported by The Sleuth Kit, including FAT16 and FAT 32, NTFS, HFS+, ext2/3/4, and ISO9660.

```

<fileobject>
  <filename>$OrphanFiles/INDEX.DAT</filename>
  <partition>1</partition>
  <id>90</id>
  <name_type>-</name_type>
  <filesize>49152</filesize>
  <unalloc>1</unalloc>
  <used>1</used>
  <inode>45990</inode>
  <meta_type>1</meta_type>
  <mode>511</mode>
  <nlink>1</nlink>
  <uid>0</uid>
  <gid>0</gid>
  <mtime>2001-07-03T06:19:54Z</mtime>
  <atime>2002-11-21T05:00:00Z</atime>
  <ctime>2001-07-02T18:41:45Z</ctime>
  <byte_runs>
    <byte_run file_offset='0' fs_offset='9715712' img_offset='9747968' len='49152'>
  </byte_runs>
  <hashdigest type='md5'>59b4d7bede8501c3ac70cd89c6184f56</hashdigest>
  <hashdigest type='sha1'>d53b5dd171246766cca64c8038d323277ffc3fee</hashdigest>
</fileobject>

```

Figure 1. Sample of XML output of fiwalk.

4.2 Bulk Data Analysis and Stream-Based Forensics

In order to identify instances of potentially private and personally identifying data, we use bulk extractor, a bulk data analysis tool also developed by Garfinkel. While there are many existing tools that can be used to parse filesystem structures and individual file formats, bulk extractor ignores file system structure and instead processes information from a disk image as a sequence of 16MiByte pages. As a practical consequence, bulk extractor can process different sections of a disk in parallel, greatly increasing performance on modern multicore hardware. Furthermore, bulk extractor can identify, decompress, and recursively reprocess compressed data (for example, modern zipped, XML-based document formats such as Office Open XML).

Bulk extractor is capable of identifying numerous data features using a collection of software modules designed to target feature types, particularly those likely to include private and personally identifying information. These include:

- Private accounts information (credit-card numbers)
- Unique private identifiers (social security numbers)
- Hexadecimal- and Base64-encoded text
- Internet domain names
- Email addresses, email messages and headers
- Ethernet MAC addresses
- EXIF metadata from JPEG images
- Telephone numbers
- URLs and search histories
- Compressed files (zip and gzip files, which BE can proactively decompress and analyse)

- Windows hibernation file fragments
- User-defined keywords

A more detailed listing can be found at https://github.com/simsong/bulk_extractor/wiki, and within the technical documentation distributed with the bulk extractor source code. A typical run of bulk extractor will produce a directory containing text files for each feature type, populated with feature instances and associated offsets into the related disk image. Histogram data for feature instances (e.g., how many times a given email address appeared on a given medium) is generated in separate files.

4.3 Data Triage

Using the filesystem information provided by fiwalk, and the feature data produced by bulk extractor, one can perform a wide variety of triage and reporting tasks working directly from an unmounted disk image. Bulk extractor includes a Python script which maps each feature instance back to a specific location on disk, allowing one to build a list of which feature instances (specifically potentially private and personally-identifying data) correspond to allocated files and which are located in deleted files or currently unallocated space.

Continuing our previous example using data from a legacy 4.3GB external USB hard disk, Table 2 illustrates that features identified by bulk extractor (in this case, URLs visited and searches performed by the user) can often be linked to extant file objects on disk even when primary application caches have been cleared.

Table 2. Example of Feature Data from Bulk Extractor

Position	16844161
Feature	http://www.yahoo.com
Context	[REDACTED@ http://www.yahoo.com \x00\xAD\x0B\...[REDACTED]
File Name	\$OrphanFiles/INDEX.DAT
File MD5	39c52b472ec890cc29f71419d6aba999

This example has been selected (and partially redacted) because it is innocuous, and cannot be linked to a specific user. However, even on this relatively small disk (notably, a disk that was never used as a primary boot medium) hundreds of examples of personally identifying and potentially sensitive feature instances were found in the linked output.

One can further identify areas where user activity and user-created data are most prevalent, by isolating collections of relevant features. For example, on a modern Windows 7 operating system this would likely include the user's home folder, the user's registry hive (NTUSER.DAT), common or shared document and media directories, hibernation files produced by the operating system, and the Recycle Bin. Bulk extractor proactively decompresses and reprocesses several common compressed formats, further reducing the need for manual intervention.

An additional method for performing data triage is fuzzy hashing. Comparing standard cryptographic hashes for two files simply indicates whether the two files are exactly the same (are the same bitstreams). Fuzzy hashes, on the other hand, can indicate whether two files are approximately

similar. This can help with the managing of private or personal data in at least two ways. First, if a user identifies an item (File A) that is relevant to his/her search goals, but File A contains private data (and thus cannot be viewed by him/her), then a repository could provide a different item (File B) that is quite similar to file A but does not contain the private data in question. Conversely, if a repository contains a set of files that are known to contain private data (based on use of a tools such as bulk extractor), then fuzzy hashing could be used to identify other similar files that might benefit from further human review to determine if they might also contain other forms of private data that were not identified in the initial investigation. BitCurator is incorporating a fuzzy hashing tool called sdhash (Roussev 2011).

4.4 Data Reporting

Using the output of fiwalk, bulk extractor, and additional open source digital forensics tools, BitCurator facilitates the creation of machine-readable and human-readable reports. These reports can be constructed in a modular fashion by reading and reprocessing Digital Forensics XML (DFXML)⁹ output and other tool-produced metadata (Garfinkel, Digital Forensics, 2012). DFXML is an evolving schema designed to simplify and standardize the interoperability of tools that produce digital forensics output. Designed to be simple to generate and reprocess, DFXML can be used to describe filesystems and the objects they contain (including low-level information such as permissions, timestamps, location on disk, and cryptographic hashes). Disk image metadata generated using DFXML-producing tools can readily be annotated with Dublin Core tags.

BitCurator will generate reports from disk image contents and DFXML output as PDFs, slide-show presentations, and machine-readable metadata. These can fill a variety of roles: highlighting distribution of data on disk (as simple charts or treemaps); listing areas likely to contain substantial amounts of private data; generating timelines of email activity; identifying use of external devices; and building a difference map between a source disk and a copy.

More sophisticated reports can be produced depending on the options one specifies for (and information one provides to) the bulk data processing tools. For example, bulk extractor supports both *stop lists* and *context-sensitive stop lists* and can be directed to suppress reporting on particular feature instances always or within certain contexts. This can be useful for organizations that are working with large volumes of material and it is known ahead of time that certain data that might appear private is, in fact, not private.

As an example, any set of feature instances that is output by the bulk extractor tool (such as a list of email addresses from a particular domain) is initially flagged with the byte offset in the disk image where it appears. Assuming the filesystem(s) on the disk have been recognized, one can run a secondary utility distributed with bulk extractor (`identify_filenames.py`) to link each feature instance either to a file currently allocated within the filesystem, a deleted file, slack within the filesystem, or an unallocated area of the disk. Subsequently, one can organize this output in a way that clearly shows the user where items of concern may appear.

⁹http://www.forensicswiki.org/wiki/Category:Digital_Forensics_XML

4.5 Redaction, Whitelisting, and Access

A significant issue with many modern digital collections is that while they may contain mostly non-private data, lack of appropriate private-data protection and distribution models often means that they remain “dark,” effectively inaccessible.

Redacted disk images can be produced by using the mapping between private data feature instances and disk blocks within the forensically-packaged disk image. Overwriting the relevant parts of the bitstream with randomized or zeroed-out data on a copy of the forensically-packaged image produces a disk image that no longer contains the private data. The redacted image can then be used in various ways—as a backing store for a web front end that provides access to contents of the filesystem, or as a complete distributable object.

Alternatively, whitelisting can be used to produce a “permissions overlay” for the original image, which may be hosted in a sandboxed virtual environment where the user (either at a local workstation or via a remote terminal) interacts with the live environment but is disallowed access to certain files or subtrees within the filesystem. This approach is particularly appropriate when the acquired medium includes a bootable operating system.

If the goal is to generate a collection known to have been produced by a single person or organization, and a donor has confirmed that all identities and data within that collection are to be made public, only a redaction profile for private data from other sources may be required. In every case, however, all private data should be clearly and systematically identified.

4.6 The BitCurator Disk Image Test Corpus

A major challenge to improving the consistency and coverage of private data handling across collecting institutions is the lack of shared corpora on which to test software designed to identify such data. It is unlikely that such a corpus could be created from existing collections materials (or unprocessed materials within backlogs) due to existing donor agreements, legal guidelines, and institutional mandates. Although it is possible to construct large corpora of data from private materials inadvertently or maliciously released onto the Web, this can be ethically problematic to further disseminate and such corpora also will not generally include the complex data structures and interrelationships found on media originally belonging to individual users and corporate entities.

Simson Garfinkel has addressed this problem in the forensics community by constructing the Real Data Corpus (RDC), approximately 30TB of data corresponding to disk images extracted from media and devices purchased on the secondary market. Researchers can obtain access to the RDC with institutional review board (IRB) approval.¹⁰ While the RDC is an extremely useful research and forensic tool development resource, it contains materials from many sources which collecting institutions would identify as having little or no archival value (for example, hard disk drives from retired ATM machines).

The National Institute of Standards and Technology has prepared a publicly available corpus, entitled the Computer Forensic Reference Data Set (CFreDS). At the time of writing, this set of disk images and exercises is largely geared towards law enforcement, and includes detailed but relatively anodyne synthesized exercises geared towards basic training and tool testing.

¹⁰ <http://digitalcorpora.org/corpora/disk-images/real-data-corpus>

For the BitCurator project, we have been constructing a test corpus in order to replicate common data analysis and triage workflow steps in digital archives. This consists of a non-public corpus of disk images extracted from fixed and removable media identified as containing data likely to have archival value (or requiring long-term preservation) by researchers and practitioners from the project's two advisory groups. In the first year of the project, we requested data from the ten project advisors on our Professional Experts Panel, and nine on our Development Advisory Group. We received data in the form of raw and forensically packaged disk images from the City of Vancouver Archives, the National Institute of Standards and Technology, Duke University, and the National Library of Australia. The transfer is based on a data transfer agreement in which the project team agrees to use the data only for purposes of research and testing within the context of the project.

We have subsequently added to this corpus approximately ten years of disk images from retired workstations and legacy external media provided by iBiblio at the University of North Carolina at Chapel Hill. Additionally, we have included approximately 100,000 government documents in common office file formats crawled from the Web for the purposes of sampling document metadata and content. The corpus currently includes approximately 7.5TB of data, with coverage of major disk formats including FAT16 and FAT32, NTFS, HFS and HFS+, ext3/4, and various double- and high-density floppy images.

In an upcoming phase of the project we will be using this corpus as a testbed for our triage, reporting, and redaction tools. This will allow us to provide statistics on consistency and coverage of our procedures, and isolate problem cases corresponding to damaged or unrecognized filesystems, rare data encodings, and any inconsistencies identified in tool output.

5. Discussion and Future Work

The ability to rapidly and accurately identify the location and nature of private data on a device has many potential applications in libraries and archives. These include:

- 1. Formal accounting of best practices—focusing on technical challenges in identifying and handling private and sensitive data in new and existing collections.**

Collecting institutions have written mandates, formal procedures, and legal guidelines in place for handling private data, but these guidelines can result in both conceptual and technical pitfalls in implementation.

- 2. Integration of existing digital forensics tools into workflows of collecting institutions.**

Tools such as bulk extractor and The Sleuth Kit provide high-quality coverage of a wide range of private information in many common disk formats, and can be extended to address the needs of those outside police investigation contexts. Ongoing and future work in this area will address the following:

- Providing cumulative statistics on what has been identified in both existing collections and raw donor materials.
- Establishing more complete chains-of-custody and technical provenance metadata in order to support records of authenticity with increased coverage and accuracy.
- Integration of existing digital forensics metadata currently used for tool interoperability (including but not limited to Digital Forensics XML) into extensible metadata schemas and standards supported by the wider community and maintained by a recognized authority.

3. Preparing customized collections for specific audiences.

Many institutions wish to provide alternate “views” of raw (unredacted) data sources based on credentials, location, and other forms of authentication. Redaction profiles that limit access to specific areas of a bitstream (or provide a complete copy of the bitstream with sensitive areas overwritten with junk data) are a natural solution to this. Providing simple links between such profiles and authentication mechanisms already in use could provide a powerful mechanism for improved public access to large backlogs of digital materials.

6. Conclusion

We have discussed basic approaches and tools for addressing private and non-private data on digital media. As part of this analysis, we have identified specific circumstances under which they can be identified and (or) redacted. We have examined some current digital forensics technologies that handle this problem efficiently with a low incidence of unwanted results. We have discussed how existing open source digital forensics tools and platforms, including The Sleuth Kit, bulk extractor, and fiwalk can be incorporated into an efficient, automated workflow in order to produce machine- and human-readable reports and metadata. We have addressed these points in the context of BitCurator, an ongoing research initiative at UNC Chapel Hill and the Maryland Institute of Technology and the Humanities to build, test, and analyse software and systems for incorporating digital forensics methods and technologies into the workflows of collecting institutions.

As stated earlier, it is important for collecting institutions to address issues of private and sensitive data for a variety of reasons: in order to serve as trusted actors who can responsibly care for digital collections, to keep the costs of processing collections from being prohibitively high, and to avoid growing backlogs of unprocessed material that are stored but not available for use. Perpetuating humanity’s heritage will increasingly involve the curation of digital traces generated by individuals (Lee 2011). Caring for and perpetuating this heritage will require responsible curation of content, attending to access restrictions and protection of individuals. Such work will require a great deal of creativity and judgment, supported by efficient and reliable tools.

Acknowledgements

This work has been supported by a grant from the Andrew W. Mellon Foundation.

References

- AIMS Working Group. “AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship.” 2012.
- Duranti, Luciana. “From Digital Diplomats to Digital Records Forensics.” *Archivaria* 68 (2009): 39-66.
- Duranti, Luciana and Barbara Endicott-Popovsky. “Digital Records Forensics: A New Science and Academic Program for Forensic Readiness.” *Journal of Digital Forensics, Security and Law* 5, no. 2 (2010): 1-12.
- Elford, Douglas, Nicholas Del Pozo, Snezana Mihajlovic, David Pearson, Gerard Clifton, and Colin Webb. “Media Matters: Developing Processes for Preserving Digital Objects on Physical Carriers

- at the National Library of Australia.” Paper presented at the 74th IFLA General Conference and Council, Québec, Canada, August 10-14, 2008.
- Garfinkel, Simson L. “AFF: A New Format for Storing Hard Drive Images.” *Communications of the ACM* 49, no. 2 (2006): 85-87.
- Garfinkel, Simson. “Digital Forensics XML and the DFXML Toolset.” *Digital Investigation* 8 (2012): 161-174.
- Garfinkel, Simson L. “Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools.” *International Journal of Digital Crime and Forensics* 1, no. 1 (2009): 1-28.
- Garfinkel, Simson. “Lessons Learned Writing Digital Forensics Tools and Managing a 30TB Digital Evidence Corpus.” *Digital Investigation* 9 (2012): S80-S89.
- Garfinkel, Simson, “Digital Forensics Research: The Next 10 Years.” DFRWS 2010, Portland, OR, August 2010
- Hibshi, Hanan, Timothy Vidas, and Lorrie Cranor. “Usability of Forensics Tools: A User Study.” In *IMF 2011: 6th International Conference on It Security Incident Management & IT Forensics: Proceedings, 10-12 May 2011, Stuttgart, Germany*, 81-91. Los Alamitos, CA: IEEE Computer Society, 2011.
- John, Jeremy Leighton. “Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools.” Paper presented at iPRES 2008: The Fifth International Conference on Preservation of Digital Objects, London, UK, September 29-30, 2008.
- Kirschenbaum, Matthew G., Richard Ovenden, and Gabriela Redwine. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Washington, DC: Council on Library and Information Resources, 2010.
- Lee, Christopher A., ed. *I, Digital: Personal Collections in the Digital Era*. Chicago, IL: Society of American Archivists, 2011.
- Lee, Christopher A., Matthew Kirschenbaum, Alexandra Chassanoff, Porter Olsen, and Kam Woods. “BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions.” *D-Lib Magazine* 18, no. 5/6 (May/June 2012). <http://www.dlib.org/dlib/may12/lee/05lee.html>
- Nesmith, Tom. “Still Fuzzy, but More Accurate: Some Thoughts on the ‘Ghosts’ of Archival Theory.” *Archivaria* 47 (1999): 136-50.
- Pearce-Moses, Richard. *A Glossary of Archival and Records Terminology*. Chicago, IL: Society of American Archivists, 2005.
- Ross, Seamus and Ann Gow. *Digital Archaeology: Rescuing Neglected and Damaged Data Resources*. London: British Library, 1999.
- Roussev, Vassil. “An Evaluation of Forensic Similarity Hashes.” *Digital Investigation* 8 (2011): S34-S41.
- Underwood, William, Marlit Hayslett, Sheila Isbell, Sandra Laib, Scott Sherrill, and Matthew Underwood. “Advanced Decision Support for Archival Processing of Presidential Electronic Records: Final Scientific and Technical Report.” Technical Report ITTL/CSITD 09-05, October 2009.
- Underwood, William, E. and Sandra L. Laib. “PERPOS: An Electronic Records Repository and Archival Processing System.” Paper presented at the International Symposium on Digital Curation (DigCCurr 2007), Chapel Hill, NC, April 18-20, 2007.

- Woods, Kam and Geoffrey Brown. "From Imaging to Access - Effective Preservation of Legacy Removable Media." In *Proceedings of Archiving 2009*, 213-18. Springfield, VA: Society for Imaging Science and Technology, 2009.
- Woods, Kam and Geoffrey Brown. "Migration Performance for Legacy Data Access." *International Journal of Digital Curation* 3, no. 2 (2008): 74-88.
- Woods, Kam and Christopher A. Lee. "Acquisition and Processing of Disk Images to Further Archival Goals." In *Proceedings of Archiving 2012*, 147-152. Springfield, VA: Society for Imaging Science and Technology, 2012.
- Woods, Kam, Christopher A. Lee, and Simson Garfinkel. "Extending Digital Repository Architectures to Support Disk Image Preservation and Access." In *JCDL '11: Proceeding of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, 57-66. New York, NY: ACM Press, 2011.
- Xie, Sherry L. "Building Foundations for Digital Records Forensics: A Comparative Study of the Concept of Reproduction in Digital Records Management and Digital Forensics." *American Archivist* 74, no. 2 (2011): 576-99.

Shared Perspectives, Common Challenges

A History of Digital Forensics & Ancestral Computing for Digital Heritage

Corinne Rogers¹ and Jeremy Leighton John²

¹*School of Library, Archival and Information Studies, The University of British Columbia, Canada, cmrogers@mail.ubc.ca;* ²*Department of Digital Scholarship, The British Library, United Kingdom*

Abstract

Although the field of computer forensics might appear distinct from the curation and preservation of cultural objects, these disciplines have overlapping histories and legacies deriving from shared challenges and theoretical perspectives. A convergence of practice of the digital forensic investigator and the digital archivist is gaining momentum as collecting institutions are faced with growing accessions of digital media. Shared theoretical perspectives include issues relating to authorship and identity, informational pattern and change over time, evidential reliability, and digital materiality. Shared challenges include the volume of a person's life information spread across myriad devices, the complexity of diverse applications and locations, the necessary versatility of tools and techniques required to capture, investigate, and describe digital information, planning to ensure sustainability and long-term preservation, and issues of security, privacy and other digital rights. This paper offers a historical overview of digital forensics mapped to current issues in digital curation and preservation in cultural heritage domains.

Authors

Corinne Rogers comes from a senior administrative background in non-profit organizations with an interest in policy and governance. Currently a doctoral student at UBC's School of Library, Archives and Information Studies, her research interests focus on legal and ethical issues associated with digital materials offered as evidence at trial. She is a research assistant under the direction of Dr. Luciana Duranti in InterPARES and the recently completed Digital Records Forensics Project. Corinne is also a researcher with the Veterans Transition Program at UBC where she is establishing an oral history project.

Dr Jeremy Leighton John has been Curator of eMANUSCRIPTS at the British Library since 2003, previously having been Specialist Scientific Curator. In 1996 he completed a Ph.D in Zoology at Merton College, University of Oxford. He is a Fellow of the Linnean Society of London and of the Royal Geographical Society. As Principal Investigator of the Digital Lives Research Project funded by the UK Arts & Humanities Research Council, he promoted archival digital forensics. Currently, a member of the Committee of the Section for Archives and Technology within the Archives and Records Association of UK & Ireland; previously, a member of the Library Committee of the Royal Society.

1. Introduction

The field of forensics and investigation might at first glance seem quite separate from the curation and preservation of cultural objects; yet these disciplines have overlapping histories and legacies deriving from similar goals, common challenges, and shared theoretical perspectives. A convergence of perspectives and methods of the digital forensic investigator and the digital archivist is gaining momentum as collecting institutions are faced with growing accessions of digital media and objects (John 2008; Kirschenbaum, Ovenden, and Redwine 2010). The conceptual underpinnings of digital (or

¹ The authors made similarly substantial contributions to the paper.

computer) forensics² can be interpreted through the lens of information and archival science, to illuminate parallels between these disciplines and reveal avenues of further research to their mutual benefit. This paper outlines points of convergence between these disciplines, offers a historical overview of digital forensics, and discusses its relevance to ancestral computing and issues of digital curation and preservation in cultural heritage domains. The role of an archivist may be as a record keeper, caring for the documentary legacy of an organization, or as a curator of a personal archive of a writer, scientist or political figure. Both functions are embraced by this paper.

2. Similar Goals

2.1 Archival Function

At the most basic level, both digital archivists and digital forensics practitioners are concerned with discovering, understanding, describing and presenting information inscribed on digital media.

The core archival functions “upon which archivists build their scientific, professional and educational profiles” (Duranti and Michetti 2012) can be identified as appraisal and acquisition, arrangement and description, retention and preservation, management and administration, and reference and access. Furthermore, research may be considered the foundation of each archival activity, “a professional function if not the core of all functions” (Duranti and Michetti 2012).

Archival research has focused historically on records, defined as documents created or received in the course of practical activity, and set aside for further action or reference (Duranti and Thibodeau 2006), as the primary objects of investigation. Records serve as evidence of actions and transactions, and lose much of their meaning in isolation or removed from their juridical or institutional context. In their analysis, the archivist strives to determine the purpose and functions of their creator, and the means and methods of their documentation. The informational content of records may in some cases be less important to this analysis than the circumstances of their creation and use. Personal archives may be investigated for their content as when a series of drafts leading to a final published novel is critically examined for evidence of literary creativity. Archivists are thus concerned with establishing the evidentiary capacity of documents, and analysing their evidential value, whether they are preserved primarily as records (as with a public organisation) or for their informational value as personal memory or legacy (as with a personal archive). According to Menne-Haritz, “Evidence means patterns of processes, aims and mandates, procedures and results, as they can be examined. It consists of signs, of signals, not primarily of words. ... All those are nonverbal signs that must be interpreted in context to disclose their meaning. To one who understands them, they will tell how processes worked and who was responsible for which decision” (Menne-Haritz 1994).

Although archival theory originated in legal and administrative doctrine reaching back many centuries to Roman times, modern archival scholarship has its roots in the 17th and 18th centuries, when it became closely aligned with historical scholarship, adopting and adapting methods of historical research and the philological disciplines to archival research into documents in the prevailing world of print.

² Early practitioners referred to the practice of computer forensics. As digital devices become ubiquitous and diverse, the term “digital” has begun to replace “computer” (see for example Whitcomb 2002) However, there is little consistency even today. While the tendency may be to preference “digital”, the term computer forensics is still in use, and is particularly appropriate in the context of ancestral computing and legacy hardware and software.

Today we live in a digital world. Computer technology has changed the way we communicate, conduct business, present our public face(s), and document our private lives. As digital communications supplant print-based culture, a new literacy is evolving. Digital culture is challenging the viability and legitimacy of many well-established social and cultural norms and their associated legal frameworks (Doueihi 2011). One aspect of this evolution can be observed in our concepts of trust in digital information and our reconception of what it means for a digital object to serve as documentary evidence, which relies on our ability to assess its authenticity and integrity. We are tempted to consider the records, documents, and information that we create and disseminate over the Internet as being equivalent to similar forms in our traditional, analogue world. Because of an assumption of functional equivalence of digital and analogue documents and data, the authenticity and trustworthiness of these new digital creations are often judged by the same standards. However, “Virtual authenticity is not to be explained by a transfer of a well-known and ultimately problematic category from one model to another; it is not to be restricted to a shift from the real to the virtual” (Doueihi 2011). Archivists are embracing the new digital literacy, re-examining traditional concepts of retained information—records, documents, data—and the attributes by which we have traditionally assessed evidentiary capacity and evidential value. Archival methodologies, developed originally over centuries in a print-based culture, are being adapted to address digital material, and are embracing new disciplinary knowledge in order to do so.

2.2 Digital Forensic Function

The new discipline of digital forensics was developed originally for the purposes of law enforcement in order to investigate computer crime and bring digital evidence to trial. It applies scientific principles and methodologies in reconstructing past events and artefacts, and has developed technologically in intimate association with the advancement of forensic tools. It is defined as:

The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence; derived from digital sources for the purpose of facilitation or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations (Palmer 2001).

Throughout its history there have been calls for digital forensics to be situated within a broader social and theoretical framework (Palmer 2001). Drawing on computer science theory and forensics theory from physical forensics disciplines, digital forensics practitioners and researchers are actively developing process models that can be used to help standardize digital forensics investigations and form the basis of new theory. These process models identify core digital forensics activities, and from these, we can draw close parallels with archival functions. Beebe and Clark (2005) have proposed one such model, and mapped their framework to similar, previous frameworks commonly cited in the literature.³ The core functions of the Beebe and Clark model are preparation and incident response, data collection, data analysis, presentation of findings, and incident closure. The first three may be compared with archival functions of appraisal and acquisition, while data analysis parallels arrangement and description, and

³ The models compared by (Beebe and Clark 2005) are: the DFRWS model (Palmer 2001; US Department of Justice 2001; Reith, Carr, and Gunsch 2002; Carrier and Spafford 2003; Mandia et al. 2003; Nelson et al. 2004, O’Ciardua in 2004; Casey and Palmer 2004).

elements of presentation of findings (also evident in arrangement and description) may be seen in reference and access, and elements of incident closure may be compared with archival retention and preservation.

2.3 Convergent Utility of Evidence

Digital forensics thus offers digital archivists another way of conceptualizing digital objects and assessing their integrity and authenticity that can complement and be complemented by existing archival methodologies (Duranti and Endicott-Popovsky 2010; Duranti and Rogers 2011). Digital forensic tools and techniques are also being applied in service of information security, including incident investigation and forensic readiness, or incident prevention (Endicott-Popovsky, Frincke, and Taylor 2007; Endicott-Popovsky and Frincke 2007; Taylor, Endicott-Popovsky, and Frincke 2007). They are increasingly being tested in trusted digital repositories to assist archival processes of acquisition, selection, appraisal, description, and preservation of cultural and scientific heritage materials (John 2008; Kirschenbaum 2008; Kirschenbaum, Ovenden, and Redwine 2010).

There is an increasing desire in both the archival and the digital forensics communities to look to knowledge in complementary disciplines to investigate the challenges to theory and practice presented by digital technologies. Social and natural sciences, humanities, law and digital forensic discipline, and the information disciplines are all evidence-based disciplines, and so have a stake in the trustworthiness of digital material and its preservation, and each brings unique theoretical perspectives to bear.

3. Common Challenges

3.1 Digital Diversity, Volume & Complexity

Some of the biggest challenges faced by digital forensic scientists and practitioners are the diversity of digital media, the sheer volume of digital material, and the complexity of the problems of capture and analysis. This similarly confronts those who are responsible for digital archives and cultural heritage repositories, for these may contain representatives of digital objects and media that come from almost anywhere. At the same time the size of the digital universe,⁴ and the rate of technological change are far outpacing the capacity of archivists and forensic examiners to keep up.

3.2 Versatility of Tools & Techniques for Digital Forensic Analysis and Curation

“Just as it has been said that ‘one software tool does not a computer examiner make,’ only possessing one investigative process model is equally as limiting. Computer forensics examiners need a repertoire of tools and just as important a repertoire of examination and investigative approaches” (Rogers et al. 2006).

To address the profusion of digital objects and media in countless situations it is necessary to have a variety of tools and not to rely on any single technology or methodology. Equally the workflows and procedures need to be receptive to change, to be designed for flexibility and adaptability so that new functions can be quickly embraced and integrated.

⁴ <http://www.emc.com/leadership/programs/digital-universe.htm>.

This novelty, diversity and complexity of tools and digital objects in turn calls for a careful and extensive marshalling of metadata and documentation. This is a requirement that may draw on the rapidly increasing experience and expertise of the digital archive and data curation communities.

Historically, much of the analysis software used by digital forensics practitioners has been proprietary and commercial or custom-built for local use. Increasingly, open source solutions are being sought and developed. This benefits forensics analysts, supports admissibility requirements of digital evidence at trial, and enhances capacity for archival processing and long-term preservation in digital repositories (Altheide and Carvey 2011).

3.3 Long Term & Lifecycle Considerations for Digital Preservation

The long-term sustainability of digital objects and the information that these bear, and their continuing accessibility and usability over time has been a primary driver, a *raison d'être*, for the digital preservation community. It is increasingly a concern for the digital forensic community too, which has made some important steps including the development of the Advanced Forensic Format, a new open source format for storing disk images (Cohen, Garfinkel, and Schatz 2009; Garfinkel 2006; Garfinkel et al. 2006; Garfinkel 2009). Although the timeframes of legal and archival activities are different, with archives generally holding material indefinitely if not 'forever', in the legal context too it is often necessary to care for artefacts for periods longer than digital media and objects can be expected to survive without due care and attention.⁵ It is an area that could benefit significantly from the attention of digital preservation developers and practitioners.

3.4 Security, Privacy, and Digital Rights

The ease with which digital material can be altered, intentionally or accidentally, and the ease with which it can be disseminated, shared, combined, and repurposed, has driven security, privacy, and rights concerns across domains and disciplines. Information and network security, measured in various types of integrity analysis and control, is foundational to digital forensics, while protection of privacy and management of digital rights are at the forefront of archival concerns. While digital forensics addresses privacy and security in the context of intrusion detection and incidence response, digital archivists and curators must be aware of privacy and rights requirements in order to manage description of and access to material entrusted to their care.

4. Shared Theoretical Perspectives

The ensuing portrayal of shared perspectives does not negate the differences in outlook and purpose between forensic examiners operating in a legal context and those adopting the techniques for the purposes of curation and preservation. In the context of law enforcement, there is a perpetrator and a victim, and digital forensics is employed to uncover and present evidence that will be material in the case at issue. Once a legal case has been completed or is closed, the evidence tends to be put aside and stored without further access or management (unless there is an appeal). Long-term preservation of supporting material, including migration to ensure continued accessibility, is often not undertaken. Scholarly

⁵ Personal communication, Tony Sammes and Brian Jenkinson, October 2011.

examination of archival material, on the other hand, continues indefinitely, and is repeated over time as new and varying research goals are adopted. The archivist who secures digital material and initially interprets it through archival analysis and description expects to be followed by historians who investigate the archive decades and centuries later. These purposes are quite different, even if the elimination of hypotheses is a critical approach in each discipline (Fraser and Williams 2011; Fraser 2010), and lead to different applications and requirements of digital forensic tools and techniques.

4.1 Authorship & Identity

There has been a longstanding interest in the identification of forgeries specifically and in ascertaining the origin of all documents generally. Forgeries were rife in medieval and classical times, requiring vigilance with respect to their authenticity at the time and their authentication subsequently by historical scholars. Examination and investigation of handwriting (palaeography) and document analysis have long played a prominent role in identifying authors (Nickell 2005).

The science of diplomatics developed since the 17th century to establish the authenticity, and indirectly, the reliability, of archival documents, in order to determine rights and to identify and eliminate forgeries. Diplomatics is concerned with proving that a document is what it purports to be through the study of its genesis, forms, and transmission, and the relationships of the documents with the actions and persons and with its juridical context. Diplomatic criticism has evolved to analyse and evaluate individual documents in terms of a recognized system of formal elements, through which those documents can be shown to have been “written according to the practice of the time and place indicated in the text, and signed with the name(s) of the person(s) competent to create them” (Duranti 1998).

In the digital environment, the absence of such elements of proof of authorship and identity as handwritten signatures have led to an erosion in trust. This has been addressed over the past 13 years in the InterPARES (International Research on Permanent Authentic Records in Electronic Systems, 1999-2012) Projects, in which the theory and methodology of traditional diplomatics and the principles of archival knowledge were refined and tested to provide a framework for assessing the authenticity of digital records through the development of a new body of knowledge, digital diplomatics. Digital diplomatics offers archivists a powerful methodology for analysing digital records. However, according to Duranti, digital diplomatics alone may not be sufficient to understand the challenges posed to information inscribed by increasingly complex digital systems (Duranti 2009).

Traditional concepts of provenance and identity are severely undermined by the default of anonymity on the Internet. The identity of creator, author, writer or originator may be obscured and separated from the inscribed message by virtue of the layers of technology that mediate between physical person and transmitted document. “Establishing who is behind the keyboard at a given time is perhaps the single most commonly litigated issue involving electronic evidence” (Scanlan 2011). Just as archival principles have been combined with diplomatic concepts to provide a methodology to analyse traditional digital objects, new knowledge from digital forensics may extend digital diplomatics to illuminate growing challenges to issues of provenance, identity, and integrity in increasingly complex digital environments.

4.2 Integrity and Change over Time

Philology, textual criticism, decipherment, stemmatics, phylogenetics and cladistics, historical relatedness, semiotics, and more recently fuzzy hashing, all share an interest in how objects change with

the passage of time. The key concept is relatedness. Historically, it has been much concerned with the development of some writing or work, and contemplates the relationship between objects, entities, existing contemporaneously (e.g., today) or longitudinally through time.

The 17th and 18th centuries saw the deepening of critical scholarship with the development of systematic investigation of literary style, poetic harmony and formal measures, chronological inconsistencies and aberrations, historical incongruities, and the presence and absence of independent, long lasting evidence (Levine 1989). The parallels in stemmatics and phylogenetics and their use in forensics, historical reconstruction, and conjectural criticism is an active area of research today (John 2009; Kraus 2009).

Integrity, once presumed from the controls on the procedures dictated by the creator of a record, now may be assessed in the absence of or further to provenance and explicit identity, and at both the physical (bits) and logical (meaning) layers of the record. Digital forensics offers another way of conceptualizing digital objects and assessing their integrity that can complement and be complemented by digital diplomacies in understanding and assessing the elements of authenticity of digital objects (Duranti and Endicott-Popovsky 2010; Duranti and Rogers 2011).

4.3 Procedures for Establishing Authenticity and Reliability of Evidence

Curation and forensics share a concern with provenance and the application of tested and certifiably reliable protocols and tools: the collection of a set of digital objects, the evidence (legal or historical), and its subsequent demonstrable authentication, once it is in the care of the responsible institution (and its evidence custodian).

Three central requirements of digital forensics match those of archivists: capturing the information without changing it, demonstrating that the information has not been changed or that the changes can be identified, and analysing and auditing the analysis of the information, again without changing it (see John 2008).

Some of the functionality of forensic software can be found in assorted tools that exist independently of the forensic community, some of it freely available on the internet. An important distinction is that forensic software is routinely subject to appraisal and peer examination (not to mention the daily scrutiny of the law courts) as well as increasingly rigorous formal testing conducted according to specified protocols and overseen by independent bodies such as the National Institute of Justice and the National Institute of Standards and Technology in the USA.⁶

Because digital evidence is extracted from digital media, its reliability and integrity depends in part on the means of its extraction, which must be conducted and accounted for according to scientific principles. The assessment of reliability and integrity is based on procedures that are repeatable, verifiable, objective, and transparent. Digital forensics tools are subject to a demonstration that they have been tested, that the error rate is known and within acceptable limits, that the tool or procedure has been published and subjected to peer review, and that it is generally accepted in the relevant scientific community (Carrier 2003). Admittedly, there is much more testing to be done and its implications remain to be comprehensively implemented and consolidated across practicing organizations.

⁶ <http://www.cftt.nist.gov/>; <http://www.nsrll.nist.gov/>.

4.4 Digital Materiality, Virtualisation & the Importance of Ornament

Just as curators and scholars attach significance to the materiality of objects, to their look and feel and behaviour, in the fullest of detail, so forensic examiners have come to realise the importance of these qualities, with for example the emergence of virtual forensic computing, involving the use of virtual machines, emulators and 3D virtual reality. This has led to the development of tools such as Virtual Forensic Computing (VFC) produced by MD5 (UK) and based on work conducted at Cranfield University in the UK (Penhallurick 2005). At the end of forensic examination and analysis it is necessary to present the material in a manner that matches (to varying degrees) the original computer environment as well as the real landscape setting itself (Schofield 2009). An example of a textbook that provides general practical advice for the presentation of digital evidence in court is that of Sammes and Jenkinson (2007).

5. Looking Back to Look Forward

5.1 The Forensics of Ancestral Computers

The rapid pace of technological change means that forensic researchers are continually having to explore new techniques and develop new tools to support security and counter criminality. Over time the modern forensics community tends to become less focused on earlier computing technologies and forensic techniques. To forensics experts concerned with staying ahead of criminal activity in cyberspace, the progress of the previous decade is becoming less relevant (Garfinkel 2010). However, the tools and techniques created to analyse older technology remain crucial to scholars such as historians and archivists, who increasingly rely on digitized and born digital material in a variety of ancestral and legacy formats. In the archival context, it is unlikely that computing in, say, the 1990s will cease to be of any interest to digital scholars. This requirement will motivate continuing research into the forensics of ancestral computers and other digital devices.

In principle, there are two major kinds of ancestral computer forensic techniques, tools and knowledge bases: (i) those developed by earlier generations of practicing computer forensic experts; and (ii) those developed subsequently as a retrospective research activity by contemporary classic computer enthusiasts, computer history conservationists, and, increasingly, digital scholars and curators. A third and emerging source of expertise might be the current generation of researchers and developers operating in the field of personal information management (PIM) and the research output of this and related fields such as software development.

Future researchers of ancestral computer forensics would benefit from being able to look back over many decades and examine the way the forensic capabilities have changed. With early disks, for example, the low track density resulted in greater space between each track, and the head would wander over the floppy disk or platter. Deletion and overwriting would be incomplete and data potentially could be retrieved from earlier writes (Nelson et al. 2004) in ways that would be more forbidding with comparable technology today.

5.2 Digital Forensics: Knowledge Retention

An important requirement from the curatorial point of view is to document digital forensics knowledge even before it ceases to be of strong relevance to the wider forensics community. The earlier experiences

of the forensics community represent an important resource, to match that of the original developers and usability designers of the hardware and software. This knowledge is already being lost, according to several forensics experts (Charters 2009; Garfinkel 2010; Pollitt 2010), and there are efforts to preserve it, at least from the perspective of the forensics community, for the purposes of better understanding future directions (Charters 2009) rather than supporting its use with obsolete media. The British Library has initiated discussions with forensic practitioners about documenting early computer forensic experiences and tools, with the aim to record interviews in due course.

5.3 History of Digital Forensics: Introduction

In the past three years three short historical retrospectives have been written that capture early development and identify future directions of digital forensics practice (Charters 2009; Pollitt 2010; Garfinkel 2010). These articles are important first-hand accounts of the evolution of the discipline and predictions for future growth reflecting the intelligence community, law enforcement and academic perspectives. Each author has been and continues to be influential in shaping the field. Each has approached the task from his particular point of view, and yet there are similarities. All accounts track the changes in computer technology, which have driven the course of digital forensics, and arrive at complementary yet distinct conclusions about future directions.

Charters' background is in IT security and information assurance spanning more than 20 years in the United States' Intelligence Community. Like Pollitt's informant, Charters believes that by "looking at the way forensics evolved in the past, with an eye to the pressures that guided its evolution, we could get a better understanding of how forensics would evolve in the near future." He explains the development of computer forensics in three stages of evolution—the Ad Hoc Phase, the Structured Phase, and the Enterprise Phase.

The twin lenses that Charters focuses on the development of digital forensics are those of policy and procedure, and forensic tool development. The Ad Hoc Phase was characterized by lack of structure, goals, Appropriate Use policy and procedures, and challenges to the accuracy of forensic tools. This phase appears to be cyclical and transcend computer technology—it happened in the mainframe era and again in the microprocessor era. (Charters wonders if it is doomed to happen again in the wake of new technologies.) The resulting confusion led to the imposition of structure expressed in policy-based programs, defined and coordinated procedures closely aligned with the policy, and a requirement for—and development of—forensically sound tools—the Structured Phase. The Enterprise Phase is characterized by real-time collection, tailored field tools and forensics-as-a-service, built seamlessly into the infrastructure. The future, he predicts, will be aimed at greater automation and interoperability, proactive collection and analysis, and increased focus on standards in software architectures and reporting.

These phases map neatly to the concept of epochs, with which Pollitt outlines the salient characteristics of the profession. Pollitt begin with "pre-history" (pre-1985 characterized by mainframe computing, an ad hoc and individualized approach to digital forensics), and, adopting a lifecycle model, moves from "infancy" (stand-alone PCs and dial-up internet, forensics in service of law enforcement, development of the first forensics groups and task forces⁷), through "childhood and adolescence" (rapid

⁷ For example, the FBI, IACIS (International Association of Computer Investigative Specialists), and IOCE (International Organization on Computer Evidence).

development of diverse technologies, broadband, increasing professionalization⁸), with “maturity” (characterized by greater opportunities for academic training, certification, standardization, and research) still to come. Within that framework he defines the discipline through the elements of people, targets, tools, organizations, and the community as a whole. Pollitt, a former military officer with over twenty years as a Special Agent of the Federal Bureau of Investigation, approaches the history through the lens of law enforcement. His experience spans the epochs he describes, and his influence is evident in the development of standards, and the recognition of digital forensics as a forensic discipline by the American Society of Crime Laboratory Directors/Laboratory Accreditation Board (Pollitt, 2010). These epochs are particularly relevant in an analysis of digital cultural heritage and the application of ancestral computing. Garfinkel, an academic practitioner who has developed computer forensics tools, conducted computer-related research and authored books and articles published in the academic and popular press, suggests a research agenda that will carry digital forensics into the next phase of development, and sets the stage by summarizing the characteristics of past phases (2010). He argues that we are coming to the end of a “Golden Age” of computer forensics characterized by relative stability of operating systems and file formats, examinations largely confined to a single computer system, removable storage devices, and reasonably good and easy-to-use tools coupled with rapid growth of research and increasing professionalism. We are facing an impending crisis, however, brought on by advances and fundamental changes in the computer industry—specifically increased storage capacity, proliferation of devices, operating systems and file formats, pervasive encryption, use of the cloud for remote processing and storage, and increasing legal challenges to search and seizure that limit the scope of investigations. This article is particularly important for its suggestion that research ought to focus on a completely new approach to understanding forensic data through the development of new abstractions for data representation. If this is supported by standards and procedures that use these abstractions for testing and validation of research products, the result will be lower costs and improved quality.

5.4 History of Digital Forensics: Organisations

The birth of the digital forensics field can be dated from the early 1980s. In 1984, the FBI established its Magnetic Media Program, which subsequently became the Computer Analysis and Response Team (CART). In 1985 Scotland Yard founded the Computer Crime Unit (CCU), which provided training courses at Interpol. The Federal Law Enforcement Training (FLETC) programs played a key role in dispensing an understanding of the recovery of digital data.

The literature is sparse for this early period in the development of investigative techniques for crimes involving computers. Most of the literature devoted to computer crime, found in computing and accounting journals, focuses on proactive security intended to prevent criminal acts by reducing opportunity (Collier and Spaul 1992). In 1986, however, an article appeared in the journal *Computers & Security* that addressed the problem from the perspective of the investigation of a crime after the fact (Stanley 1986).

Possibly the first published use of the term “computer forensics” in the academic literature appeared six years later in an article entitled “A forensic methodology for countering computer crime”

⁸ Including establishment of SWGDE (Scientific Working Group on Digital Evidence), DFRWS (Digital Forensic Research Workshop), and IFIP (International Federation for Information Processing) Working Group on Digital Forensics.

(Collier and Spaul 1992). Collier and Spaul proposed the term “computer forensics” as a label for “existing but very limited activities amongst the police and consultancy firms” and advocated for its inclusion in the realm of traditional forensic sciences. They identified the skills required of a computer forensic expert to be multi-disciplinary, including investigative capacity, legal knowledge (including the law of evidence, rules of hearsay and admissibility), courtroom presentation skills as well as knowledge of computers. Although the term appears not to have existed formally in print prior to this publication, it had long been used informally (Sommer 1998; see also Sommer 1992)

The bulk of published material begins in the mid-1990s, originating from international gatherings of law enforcement. Some of these, like the FBI international conferences on computer evidence, were symposia devoted to computer crime (Noblett, Pollitt, and Presley 2000). Others were long-established gatherings that began to include sessions on computer forensics, such as Interpol’s International Forensic Science Symposia.

It was not until the 1990s that standardisation and dissemination began in earnest. In the UK the Computer Misuse Act was introduced in 1990, and was specifically designed to deal with criminal activities associated with computer systems, which in turn motivated greater attention towards evidence handling and other procedures.

The Royal College of Military Science (later the Defence Academy of the UK) commenced courses soon afterwards: initially as short courses, subsequently at MSc and Ph.D. levels. The Interpol European Working Party on Information Technology Crime was established in 1993, and quickly followed by the formation of the International Organisation on Computer Evidence in 1995.

Computer forensics has been defined as “the application of science and engineering to the legal problem of digital evidence,” and the admissibility of digital evidence has been related to the legal requirements of comparable paper-based evidence (Pollitt 1995). Pollitt compares the document paradigm (traditional, paper-based) with the (then) new digital paradigm in order to situate digital evidence within the legal process. He summarizes the investigative process for traditional document in four phases: acquisition -> identification -> evaluation -> admission of evidence. This has become the basis for all subsequent process models.

A handbook entitled *Computer Evidence*, produced by Edward Wilding, 1997, outlined techniques for PCs using DOS and the use of DIBS (Digital Image Backup System) and included details about the handling of evidence (Wilding 1997). This was followed by *Forensic Computing*, by Tony Sammes and Brian Jenkinson (2000), which combined first principles with practical investigation and incorporated a pioneering section on electronic organisers (a second edition appeared in 2007). It contains some of the original source references used by the forensic community at the time. Other influential texts such as *File System Forensic Analysis* by Brian Carrier and *Forensic Discovery* by Dan Farmer and Wietse Venema, both published in 2005, are listed at the end of this paper (Carrier 2005; Farmer and Venema 2005).

5.5 History of Computer Forensics: Software

Tools were produced by government agencies such as the Royal Canadian Mounted Police (RCMP) in Ottawa and the US Internal Revenue Service (IRS); but over the years, many advances in the field have been due to individual law enforcement officers and to computer technicians in the private sector (Sommer 1998).

Some investigators wrote their own tools in C and assembly language; examiners carried out their work from the DOS command prompt and through the use of hexadecimal editors. The Windows environment was avoided due to its propensity to alter data and to write to the drive.

A tool called X-Tree Gold arose in the 1980s, useful for recognising file types and retrieving lost or deleted files. Norton Disk Edit soon followed, and became a preferred tool. Later still came an expanded set of Norton Utilities as well as PC Tools. Of distinctive interest are the GammaTech Utilities for inspection and data recovery of computers with OS/2 (Wilding 1997; Nelson et al. 2004).

In the 1990s prominent suites of command line utilities became available: Maresware Forensic Processing Software Suite; and The Coroner's Toolkit (TCT). Designed for the forensic analysis of compromised Unix systems, TCT suffered from not being portable between systems and from its inability to cater for file systems other than those of Unix. It has since evolved into Sleuth Kit and Autopsy. Also noteworthy from this period were DriveSpy, Image and PDBlock of Digital Intelligence.

Around this time, ASR Data created Expert Witness for the Macintosh, the GUI forerunner of EnCase. The two GUI tools that would come to dominate computer forensics both arose in the late 1990s: EnCase of Guidance Software and Forensic Tool Kit of AccessData. For the purpose of viewing and handling a wide variety of file types, it was common for forensics experts to adopt existing viewing software such as QuickView Plus, Outside In, LView Pro, Magellan, ACDSee, ThumbsPlus and IrfanView.

A brief example of how software evolves, and how one developer picks up where another left off, is provided by file carvers which assist in the recovery of deleted files, acting independently of the file system by simply seeking to identify files through their file signatures and other identifiers. The carving tool Foremost was originally created in March 2001 by Jesse Kornblum and Kris Kendall from the United States Air Force Office of Special Investigations for analysing and recovering deleted files. It was itself inspired by the program called CarvThis that had been developed in 1999 by Defense Computer Forensics Lab although it was never released to the general public. Foremost is now open source, and the code is maintained by Nick Mikus (Spenneberg 2008).

Another tool known as Scalpel was derived from Foremost 0.69 by Golden G. Richard III, and Scalpel became highly regarded, said to be recommended by Foremost developers themselves. In recent years, Foremost has received yet more attention: "Although Scalpel was far superior to its predecessor in 2005—with the ability to analyse images around 10 times faster—Foremost has caught up recently thanks to Nick Mikus, and it is actually superior to its derivative for some tasks" (Spenneberg 2008).

5.6 History of Computer Forensics: Legacy Machines

Nelson and colleagues have recommended that forensic practitioners retain legacy equipment for as long as possible — both software and hardware — suggesting that there are some forensic tasks using current operating systems and hardware, that cannot be performed with modern tools: "To be an effective computing investigator and forensic examiner, you should maintain a library of old operating systems and application software" (Nelson et al. 2004).

Even as the more universal and forensically dedicated tools were becoming firmly established, practitioners were reminding colleagues of the need to retain earlier computers and tools, as Microsoft, Apple and Unix/Linux became predominant. Computers such as Commodore 64, Osborne I or Kaypro running CP/M may be needed; or even a Pentium I or 486 PC for accessing an early IDE disk (Nelson et al. 2004).

5.7 Data Recovery

Alongside the development of security and forensic computing, other professions were already conducting many of the activities that are today embraced by forensics (Kane and Hopkins 1993, with a 3.5 inch floppy disk that “Contains 16 life saving utilities”). There is a multitude of publications on data recovery each catering for the diverse early computer systems as well as current ones. Data recovery companies today include: MjM data recovery, Disklabs, Xytron Data Recovery, DPTS, eMag Solutions). Professional associations include: Global Data Recovery Alliance;⁹ and International Professional Data Recovery Association (IPDRA).¹⁰

5.8 Reverse Engineering for Interoperability

Reverse engineering was being conducted not only by those seeking to understand software or computer programs (including malware) (Eilam 2005) but in order to understand file formats and so be able to address interoperability on behalf of publishers. For example, InterMedia (UK) produced a floppy disk conversion system so that files on an extensive variety of disks could be copied to a publisher’s own editing system (see John 2008). A longstanding tool in reverse engineering continuing today is IDA, an interactive disassembler and debugger¹¹ (Eagle 2008).

6. Classic, Retro, Vintage Computing & Gaming

6.1 Rosetta Machines

Digital scholars have invoked the concept of the Rosetta machine, any computer that is peculiarly preadapted for the interflow of information between generations of computers and storage media, citing the Macintosh Wallstreet edition of the G3 PowerBook as the holotype for this kind of technology: manufactured in 1998 it came with swappable CD, DVD, and floppy drives capable of reading 800K and 400K disks, an Ethernet port, a PCMCIA slot permitting the addition of USB ports and access to an external hard drive; a swappable zip drive was also available (Kirschenbaum, Ovenden, and Redwine 2010).

A possible candidate as a Rosetta machine is the Cat Mac, a system that was designed to allow individuals to build their own Apple Macintosh computers (Brant 1991). The guide to the system also highlights the possibility of running DOS on a Mac, and an Apple OS on a PC or Atari. In addition to matching specific models it was possible to mix and match some of the components; and it might in this way be possible to build a number of dedicated and versatile Rosetta machines—if the necessary components could be found. It accepted hard drives including a removable version by Wetex, floppy drives, Iomega’s Bernoulli design (predecessor of the Zip disk), SyQuest’s and Ricoh’s removable cartridges, optical drives (Magneto Optical, MO; Erasable Optical, EO; Write Once Read Many, WORM; Read Only, CD-ROM); tape drives including data cassettes, akin to audio cassettes but significantly different internally, and DAT 4mm cartridges; limited networking is also feasible.

⁹ <http://www.globaldra.org>.

¹⁰ <http://ipdra.org>.

¹¹ <http://www.hex-rays.com/products/ida/overview.shtml>.

6.2 Enthusiast Computing

Due to their ‘universal computer’ nature, computers lend themselves to self design and manipulation. Personal computing was famously driven by hobbyists experimenting at home. Apple I was a circuit design and board rather than a complete computer and even with later finished models, enthusiasts were invited to build their own—a tradition that continues to this day (Owad 2005). Subsequently this was discouraged by Apple but remained commonplace with PCs. Customers were encouraged to upgrade their computers themselves, and often needed to install their own drivers and generally tinker.

This ‘universality’ led, of course, to numerous types of computing devices; but it also yielded a large community of enthusiasts, many of them associated with games and gaming: in time, gradually evolving into a classic and retro computing community—with much of the activity being directed at the capture, emulation and preservation of computer games (John 2008; McDonough et al. 2010, see also recent news concerning the Princess of Persia source code).¹² The role of ancestral computing enthusiasts has been outlined previously (John 2008); further examples are a web site dedicated to the Spectrum computer¹³ and another that is concerned with the capture of Apple II disks (with a link to a video on how to clean the disk drive).¹⁴

Computer experts with an interest in games and other early computer technologies specialised in this area and made it possible to capture and interpret a wide variety of floppy disk formats, notably Disk2FDI which produces disk images in an open file format and has been used to process a major collection of games for the BnF, Bibliothèque nationale de France.¹⁵

6.3 Specialist Modern Technology

As Doug Reside has observed, ancestral (obsolete) equipment, such as a computer with a 5.25” floppy disk drive, requires considerable attention and care; and in the long run modern specialist replacement technology is the sustainable option (Reside 2010; Reside 2011).

The Software Preservation Society,¹⁶ a group derived from an interest in gaming and initially the Amiga computer, has made available the KryoFlux and accompanying software, a device for reading floppy disk drives via a USB connector. It works as standard with 3.5” and 5.25” floppy disks and has been made to work recently¹⁷ with 8” floppy disks with the help of the FDADAP floppy disk adapter.¹⁸ The recently formed company has been exploring a number of business models, selling the basic setup to individuals at a lower price while charging institutions a higher fee, with support provided. There is a GUI emerging as an advanced form of DTC (DiskTool Console), and a forensic version of the KryoFlux has been promoted.

An important consideration is the nature of the information that the KryoFlux yields; there are detailed log files that document the capture with hash values calculated. There are two kinds of capture

¹² <http://www.wired.co.uk/news/archive/2012-04/23/prince-of-persia-source-code>, and <https://github.com/jmechner/Prince-of-Persia-Apple-II>.

¹³ <http://www.worldofspectrum.org/emulators.html>; and <http://www.worldofspectrum.org/TZXGuide.pdf> for converting tapes.

¹⁴ <http://adtpro.sourceforge.net/>.

¹⁵ <http://www.joguin.com>; <http://www.oldschool.org/disk2fdi/>.

¹⁶ <http://www.softpres.org>.

¹⁷ <http://forum.kryoflux.com/viewtopic.php?f=3&t=159>.

¹⁸ <http://www.dbit.com/fdadap.html>.

output: fundamentally the stream files which represent the magnetic flux transitions; and interpreted sector disk images. The stream files are not intended for long term preservation as their encoding is optimised for the transfer of the binary stream “over the wire while a track is being read.” Software Preservation Society state: “Long term storage of disk data should be kept in the DRAFT format.” However, DRAFT (Data Recording Archival Flux Format) is not yet complete.¹⁹ There is another format known as IPF, Interchangeable Preservation Format, and the forum for KryoFlux indicates that the company has long reflected on how best to licence the IPF library source code; it has released the IPF decoder library under a modified MAME licence.²⁰

Visualisation tends to be seen as a tool that is useful at the end of the curatorial or forensic lifecycle but it can also be used at other stages. A new software tool for use with the KryoFlux demonstrates the value of visualising the condition and formatting of floppy disks at the level of the magnetic flux transitions: in identifying unformatted, healthy and degraded disks and much else besides. For further details see the KryoFlux website.

Other useful tools for working with floppy disks include the Catweasel, Disc Ferret and FC5025 controller.²¹

6.4 Reconfigurable Hardware, Emulation Software

Despite these important advances, digital capture is still dependent on the original disk drives. Perhaps an enterprising group will create a modern version of disk drives, using reconfigurable hardware. A step in the right direction is the C-One which is a modern computer (motherboard) designed to be reconfigured to behave like one or more earlier computers.²²

The ultimate in reconfigurability, of course, lies in the virtualisation of machines and components. Much of the understanding and the techniques being developed in the design of emulators and universal virtual machines are truly forensic in approach and standard. Important projects in this area include Dioscuri, Qemu, KEEP, POCOS and Planets.²³

7. Software Licensing & Preservation

The curation of personal archives where the original look and feel are critical for scholarship currently relies on the use of ancestral software. Most especially, the booting from an original disk image entails the use of the original software residing on the disk. What are the implications for curatorial examination, for access, for long-term preservation? (In July 2012, The Court of Justice of the European Union ruled

¹⁹ <http://www.softpres.org/kryoflux:stream>; <http://www.softpres.org/glossary:draft>.

²⁰ <http://forum.kryoflux.com/viewtopic.php?f=2&t=136>; <http://forum.kryoflux.com/viewtopic.php?f=2&t=265>; <http://en.wikipedia.org/wiki/MAME#License>.

²¹ <http://www.jschoenfeld.com/home/indexe.htm>; <http://discferret.com/wiki/DiscFerret>; <http://mith.umd.edu/vintage-computers/fc5025-operation-instructions>; <http://www.deviceside.com/>.

²² <http://www.c64upgra.de/c-one>; for a brief introduction see (John 2008).

²³ <http://dioscuri.sourceforge.net/dioscuri.html>; http://wiki.qemu.org/Main_Page; <http://www.keep-project.eu/ezpub2/index.php>; <http://www.pocos.org/index.php/pocos-symposia>; <http://planets-suite.sourceforge.net/opf/>.

that it is legal to trade ‘used’ software provided the reseller makes his or her own copy unusable following the sale, a decision that seems to have potentially beneficial implications for digital preservation.²⁴⁾

A prototype, somewhat procrustean, policy might be that the originators agree to forego further use of the software so that the curators may inherit the right to use it on behalf of the repository institution. In fact software licensing is extremely diverse and complex. Some of the issues of software preservation generally have received attention in recent years (Matthews et al. 2008), not least in the context of computer games and virtual worlds (McDonough et al. 2010; see also John et al. 2010) A Technology Watch paper from the Digital Preservation Coalition by Andrew Charlesworth has examined Intellectual Property.²⁵

8. Lessons for Mobile Forensics?

Challenges in the forensics of handheld and mobile devices arise primarily from the diversity of approaches taken by systems with many idiosyncratic qualities, which differ from those that are familiar from desktop and laptop forensics: platforms, encodings, operating systems, memory systems, interface methods, storage technologies, software agents and APIs (Application Programming Interface), timestamps all may differ in special and frequently proprietary ways; and products from the same vendor can differ significantly. It is reminiscent of the early days of personal computing except that mobile devices change even more frequently. Exemplars of the devices as well as a stockpile of power and data connectors and cables will be invaluable in the years ahead.

Is it possible that experiences with early computing can provide lessons for the present in this respect?

9. Tools & Software Repository

Planets²⁶ initiated a registry of software tools suitable for digital preservation, and the Digital Lives Synthesis promoted the notion of corpora for the digital curation of personal digital objects that can serve for archival testing. In addition to collecting software, ancestral software as a matter of urgency, it would be beneficial for some centres to collect forensic tools and software, especially those used during the early history of computer forensics.

There are tools for changing metadata (inappropriately) such as date-time; but these have the (fortunate) weakness that few are truly certified in their ability to do what they claim to do and may be incomplete in their actions. On the other hand it means that an understanding and documentation of just what these tools do may be necessary in order to detect forgery.

²⁴ see Press Release Number 94/12, Luxembourg, 3 July 2012, <http://curia.europa.eu/jcms/upload/docs/application/pdf/2012-07/cp12009en.pdf>.

²⁵ See <http://www.dpconline.org/advice/technology-watch-reports>. Member login required at present but will be available by time of publication of this paper.

²⁶ The successor of the digital preservation project Planets which was originally funded by the European Union is <http://www.openplanetsfoundation.org/>.

10. Knowledge Reuse & Documentation

The field of digital forensics may be young, but it is rich and complex. From its beginnings as a technical support to law enforcement, it has developed into an accepted forensic science, and its tools and techniques are valuable in both reactive and proactive and preventative endeavors. At least within the close confines of the legal forensics community, there is an emerging tradition of case files, anecdotes and instances from which to draw experience and judicious insights. Curatorial and archival forensics and digital scholarship have only just begun in the field of forensic and ancestral computing.

Documentation is a persistent problem. As Linus Torvalds replied on being asked—around the time when Linux 3.0 became available—to state the toughest technical problem during the development of Linux so far: “The two areas where we’ve had serious problems was documentation and help from hardware manufacturers.”²⁷

A potential area of collaboration among institutions and across forensic and preservation disciplines is in the archiving of manuals, datasheets, for hardware as well as software, and in approaching computer companies. Ronald van der Knijf warns that existing embedded systems will have been replaced within a decade: “There is a high demand for cooperation with the industry because a lot of time is spent building knowledge about the working and behavior of systems that are designed and built by people who already have most of that knowledge but are not allowed to share it” (van der Knijff 2012).

11. Virtualisation of Ancestral Computers

The virtual experience is highly portable, meaning that it can be made available on workstations with restricted access, a single reading room or, where appropriate and permitted, on the web. The usual way to interact with the original system is to ‘restore’ the original disk (as a clone, a hard drive) from the disk image, and boot up this clone. An alternative, more flexible, approach is to make use of virtual computing tools whereby virtual hard drives are accessed by a virtual machine (Barrett and Kipper 2010)

The possibility of using a disk image to virtually boot disks derived from personal archives was discussed in John (2008). The initial but still evolving approach at the eMSS Lab in the British Library may be outlined. The software Mount Image Pro of GetData²⁸ may be used to mount multiple ‘dd’ disk image files as well as virtual file systems such as those of VMware; it also works with the AFF image file and the proprietary one of EnCase. Having mounted the disk image as an emulated physical disk, a suitable virtual machine can be created (e.g., VMware) and the ‘raw disk’ can be added to the virtual machine, which is booted up using the original operating system that resides in the ‘raw disk’.²⁹ Virtual machines may be configured manually or quasi-automatically although some tweaking is often necessary.

One of the key advantages of using a virtual environment (instead of booting a physical clone of a physical disk) is that it allows for highly repeatable, controlled and referable experiences suitable for the scholar or scientist who is required to study the subject in an academically accountable way.

²⁷ Interview of Linus Torvalds, *Linux Magazine* 131 (October 2011): 16-18.

²⁸ <http://www.getdata.com>; <http://www.mountimage.com>; other useful tools include LiveView, VirtualBox, Xen, VMware and VMware Fusion, and Parallels.

²⁹ Sometimes it is helpful to first create a virtual disk (e.g., using VMware) and then to clone the ‘raw disk’ to this virtual disk (see Penhallurick 2005).

A common approach is to build a virtual environment from scratch: first the operating system, then the necessary application software and then add the files for viewing. Frequently it requires careful configuration and testing. A significant advantage of capturing entire disks—or at least capturing the software (and profiles of the hardware and services) along with the focal files is that the captured disk ‘image’ comes readymade (in many cases) with the appropriate settings and preferences of the archive’s originator, notably with the application software satisfactorily tuned with the operating system; moreover, the captured system is authenticated by means of the hash values, and software can be identified in detail by means of hash libraries.

It enables scholars to explore and immerse themselves in an environment that matches that of the creator, with a virtualised equivalent of the original folder structure and computer desktop arrangement in place and files available for examination. It is even possible, depending on availability and condition to set up this environment on the original computer with a new hard drive. Where a person retains system snapshots (or shadow volumes) of the entire contents of a computer throughout its life, it will be possible for the researcher to follow its evolution through time.

Sometimes the originator’s setup will be awry and the system prone to malfunction; in which case some accountable tinkering or in depth manipulation of a replicate of the archival disk image may be necessary to produce an access version of the system.

Another issue is the speed of response under emulation: for some scholarly purposes it may be too slow. A possible solution would be to offer an option for a realistic pace and behaviour, and another option for a more responsive interaction, depending on the requirements of the scholar. An ultimate goal would be to engineer a combination of high performance searching and exploration with the option to switch where desired to high fidelity viewing.

Scholarly note taking may be enhanced through the retention of VM snapshots or through video capture of screen activity. A challenge remains in devising a widely accepted means of referencing the virtual experience for academic research.

12. Open Source Emulators of Ancestral Computers

It is possible to use emulators of ancestral computers for the same purpose. For example, the open source emulator SheepShaver has been used by the British Library to mount and boot a forensically sound disk image of a hard drive from a G3 PowerPC Apple Macintosh with OS 8.³⁰ The computer in question is from the archive of the evolutionary biologist W. D. Hamilton. Each time the disk image is booted within the SheepShaver emulator, one of a number of desktop pictures (e.g., the forest and river system of Amazonia) becomes apparent along with the numerous ‘objects’ on the virtual desktop; if the emulated computer is left inactive for a while, a screen saver from the original computer appears. Files can be opened using the original software that resides on the G3 hard drive including Microsoft Word 4, Acrobat Reader and Photoshop 4. Most excitingly, it is possible to run C++ programs using the CodeWarrior software that exists on the original disk, with dynamical graphs of the program output; moreover, users may potentially change the parameters of the original program and run a modified version for themselves.

The archival technical team at Emory University has demonstrated a similar arrangement for the digital archive of Salman Rushdie, and as Naomi Nelson observed:

³⁰ <http://www.emaculation.com/doku.php/sheepshaver>.

The emulated environment is as close as we can get to recreating Rushdie's desktop for the researcher. Instead of isolated files on floppies, we have the entire context in which he worked. We can see how he used technology and the growing Web as part of his creative process.³¹

13. Reservation, Redaction, Confinement

Tools for manipulating disk images and virtual machines are useful in several ways but most critically for editing disk images in order to comply with privacy requirements; some files may be embargoed for a number of years or decades. Following selection of digital objects that have been agreed and accepted for accession by the repository, Winimage, for instance, can be used to delete folders and files from access versions to cater for privacy agreements while retaining bootable functionality. An important aspect for future research is for the security of this editing and deletion process to be accountable and demonstrably secure.

14. Conclusion

Descriptive process models, however generalized, are necessarily limited in their ability to suggest a theory of digital forensics that identifies concepts and functional requirements of the discipline. Beebe and Clark hint at this in their articulation of digital investigative principles (Beebe and Clark 2005). The goal is to develop a conceptual model that is based on more than "investigative experiences and biases" (Carrier and Spafford 2006). A model that succeeds in this will conceptualize the requirements for "forensic soundness" of the resulting analysis and support the development of procedural methods and tools (Casey 2007).

At the most basic level, abstracted models of digital forensics activities are based on three major functions: acquisition, analysis, and presentation (Carrier 2003). The theoretical concepts to be embedded in and realized through any abstract model are revealed in principles of practice that protect the authenticity, integrity, and reliability of digital material for the immediate purpose, whether that be presentation of digital evidence at trial or investigation of intrusion. Increasingly, a fourth function could be proposed—that of long-term preservation as required, or at least the possibility of long-term preservation.

This can be compared with key archival functions, articulated in a wealth of scholarly archival literature, and central to the work of digital heritage preservation: appraisal and acquisition, arrangement and description, retention and preservation. Preservation of digital heritage for use by current and future scholars depends on tools and methodologies that protect the authenticity, integrity, and reliability of digital material, and ensure its accessibility and usability over time and across technological change. Well-established for documents and records in traditional media, the affordances of digital technologies are forcing archivists to take up new tools and techniques from the digital forensics toolkit.³²

The future, for cultural heritage, depends on digital forensics knowledge from the past.

³¹ Naomi Nelson, <http://web.library.emory.edu/node/358>.

³² There is a pending Digital Preservation Coalition Technology Watch report by Jeremy Leighton John entitled: Digital Forensics and Preservation.

References

- Altheide, Cory, and Harlan Carvey. *Digital Forensics with Open Source Tools*, 1st ed. Syngress, 2011.
- Barrett, Diane, and Greg Kipper. *Virtualization and Forensics: A Digital Forensic Investigator's Guide to Virtual Environments*, 1st ed. Syngress, 2010.
- Beebe, Nicole Lang, and Jan Guynes Clark. "A Hierarchical, Objectives-based Framework for the Digital Investigations Process." *Digital Investigation* 2, no. 2 (2005): 147–167.
<http://www.sciencedirect.com/science/article/B7CW4-4G7JXTM-1/2/aa0ba8969e75dd766b77a09b1e38967e>.
- Brant, Bob. *Build Your Own Macintosh and Save a Bundle*. Windcrest Books, 1991.
- Carrier, Brian. "Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layers" *International Journal of Digital Evidence* 1, no. 4 (2003): 1–12.
- . *File System Forensic Analysis*, 1st ed. Addison-Wesley Professional, 2005.
- Carrier, Brian, and Eugene Spafford. "Getting Physical with the Digital Investigation Process." *International Journal of Digital Evidence* 2, no. 2 (2003): 1–20.
- Carrier, Brian, and Eugene H. Spafford. "Categories of Digital Investigation Analysis Techniques Based on the Computer History Model." *Digital Investigation* 3, suppl. 1 (2006): 121–130.
<http://www.sciencedirect.com/science/article/B7CW4-4KCPVBY-5/2/affc23432b1b89386dda701e2b554791>.
- Casey, Eoghan. "What Does 'forensically Sound' Really Mean?" *Digital Investigation* 4, no. 2 (2007): 49–50. <http://www.sciencedirect.com/science/article/B7CW4-4NWNCS-1/2/36717bc8a1dc225cfec6a4c835866999>.
- Charters, Ian. (2009). "The Evolution of Digital Forensics: Civilizing the Cyber Frontier." <http://www.guerilla-ciso.com/wp-content/uploads/2009/01/the-evolution-of-digital-forensics-ian-charters.pdf>.
- Cohen, Michael, Simson Garfinkel, and Bradley Schatz. "Extending the Advanced Forensic Format to Accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information and Forensic Workflow." *Digital Investigation* 6(Supplement 1) (2009): S57–S68.
<http://www.sciencedirect.com/science/article/B7CW4-4X1HY5C-9/2/5c0d842b6ee8802cfe11230246fc9772>.
- Collier, P. A., and B. J. Spaul. "A Forensic Methodology for Countering Computer Crime." *Artificial Intelligence Review* 6 (1992): 203–215.
- Doueih, Milad. *Digital Cultures*. Cambridge, MA: Harvard University Press, 2011.
- Duranti, Luciana. *Diplomatics: New Uses for an Old Science*. Lanham: Scarecrow Press, 1998.
- . "From Digital Diplomatics to Digital Records Forensics." *Archivaria* 68 (Fall 2009): 39–66.
- Duranti, Luciana, and Barbara Endicott-Popovsky. "Digital Records Forensics: A New Science and Academic Program for Forensic Readiness." *Journal of Digital Forensics, Security and Law* 5, no. 2 (2010): 1–12. <http://www.jdfsl.org/subscriptions/JDFSL-V5N2-Duranti.pdf>.
- Duranti, Luciana, and Giovanni Michetti. "Archival Method." Vancouver, BC, 2012.
- Duranti, Luciana, and Corinne Rogers. "Educating for Trust." *Archival Science* 11(3-4) (2011): 373–390. doi:10.1007/s10502-011-9152-3. <http://www.springerlink.com/index/10.1007/s10502-011-9152-3>.

- Duranti, Luciana, and Kenneth Thibodeau. "The Concept of Record in Interactive, Experiential and Dynamic Environments: The View of InterPARES." *Archival Science* 6, no. 1 (2006): 13–68.
- Eagle, Chris. *The IDA Pro Book: The Unofficial Guide to the World's Most Popular Disassembler*. No Starch Press, 2008.
- Eilam, Eldad. *Reversing: Secrets of Reverse Engineering*, 1st ed. Wiley, 2005.
- Endicott-Popovsky, Barbara, and Deborah Frincke. "Embedding Hercule Poirot in Networks: Addressing Inefficiencies in Digital Forensic Investigations." In *Foundations of Augmented Cognition, Lecture Notes in Computer Science* 4565 (2007): 364–372. http://dx.doi.org/10.1007/978-3-540-73216-7_41.
- Endicott-Popovsky, Barbara, Deborah A. Frincke, and Carol Taylor. "A Theoretical Framework for Organizational Forensic Readiness." *Journal of Computers* 2, no. 3 (2007): 1–11. www.academypublisher.com/ojs/index.php/jcp/article/download/.../291.
- Farmer, Dan, and Wietse Venema. *Forensic Discovery*, 1st ed. Addison-Wesley Professional, 2005.
- Fraser, Jim. *Forensic Science: A Very Short Introduction*. Oxford University Press, 2010.
- Fraser, Jim, and Robin Williams, eds. *Handbook of Forensic Science*. Willan, 2011.
- Garfinkel, Simson L. "AFF: A New Format for Storing Hard Drive Images." *Communications of the ACM* 49, no. 2 (February 2006): 85–87. <http://vnweb.hwwilsonweb.com/hww/jumpstart.jhtml?recid=0bc05f7a67b1790e939d3f3af5fcffbb4fe36e4feca9cd3be8b83614f440a32063db8c63b3ca06c6&fmt=C>.
- . "Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools." *International Journal of Digital Crime and Forensics* 1, no. 1 (2009): 1–28. <http://simson.net/clips/academic/2009.IJDCF.AFFLIB.pdf>.
- . "Digital Forensics Research: The Next 10 Years." *Digital Investigation* 7 (2010): 64–73. doi:doi.101016/j.diin.2010.05.009. www.elsevier.com/locate/diin.
- Garfinkel, Simson L., K. Malan, C. Dubec, C. Stevens, and C. Pham. "Advanced Forensic Format: An Open, Extensible Format for Disk Imaging." In *Research Advances in Digital Forensics Second Annual IFIPWG 11.9 International Conference on Digital Forensics*. Springer, 2006.
- John, J.L. (2008). "Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools." In *The Fifth International Conference on the Preservation of Digital Objects (iPRES)*.
- . "The Future of Saving the Past." *Nature* 459 (June 11, 2009): 775–776.
- John, J. L., I. Rowlands, P. Williams, and K. Dean. *Digital Lives: Personal Digital Archives for the 21st Century - An Initial Synthesis*. Digital Lives Research Paper, 2010, <http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf>.
- Kane, Pamela, and Andy Hopkins. *The Data Recovery Bible, Preventing and Surviving Computer Crashes/Book and Disk*. Brady: Pap/Dis, 1993.
- Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination*. Cambridge, MA: MIT Press, 2008.
- Kirschenbaum, Matthew G., Richard Ovenden, and Gabriela Redwine. *Digital Forensics in Born Digital Cultural Heritage Collections*. Washington, D.C.: Council on Library and Information resources, 2010.
- van der Knijff, R. (2012). "Embedded Systems Analysis." In *Handbook of Digital Forensics and Investigation*, edited by Eoghan Casey. London: Elsevier Academic Press.

- Kraus, Kari. "Conjectural Criticism: Computing Past and Future Texts." *Digital Humanities Quarterly* 3, no. 4 (2009). <http://digitalhumanities.org/dhq/vol/3/4/000069/000069.html>.
- Levine, J. "'Et Tu Brute?' History and Forgery in 18th-century England." In *Fakes and Frauds: Varieties in Deception in Print and Manuscript*, edited by R Myers and M Harris. Winchester: St. Paul's Biographies, 1989.
- Matthews, Brian, Brian McIlwrath, David Giaretta, and Esther Conway. *The Significant Properties of Software: A Study*. JISC, 2008.
- McDonough, Jerome P., Robert Olendorf, Matthew Kirschenbaum, Kari Kraus, Doug Reside, Rachel Donahue, Andrew Phelps, Christopher Egert, Henry Lowood, and Susan Rojo. "Preserving Virtual Worlds Final Report." 2010. <http://hdl.handle.net/2142/17097>.
- Menne-Haritz, Angelika. "Appraisal or Documentation: Can We Appraise Archives by Selecting Content?" *The American Archivist* 57, no. 3 (July 1994): 528–542. <http://www.jstor.org/stable/40293851>.
- Nelson, Bill, Amelia Phillips, Christopher Steuart, and F. Enfinger. *Computer Forensics and Investigations*. Boston, MA: Thomson Learning, 2004.
- Nickell, Joe. *Detecting Forgery: Forensic Investigation of Documents*. The University Press of Kentucky, 2005.
- Noblett, M.G., Mark M. Pollitt, and L.A. Presley. "Recovering and Examining Computer Forensic Evidence." *Forensic Science Communications* 2, no. 4 (2000): 1-13. <http://www.fbi.gov/hq/lab/fsc>.
- Owad, Tom. *Apple I Replica Creation: Back to the Garage*, 1st ed. Syngress, 2005.
- Palmer, G. "A Road Map for Digital Forensic Research." DFRWS Technical Report, 2001. <http://www.dfrws.org/2001/dfrws-rm-final.pdf>.
- Penhallurick, M.A. "Methodologies for the Use of VMware to Boot Cloned/mounted Subject Hard Disk Images." *Digital Investigation* 2 (2005): 209–222.
- Pollitt, Mark M. "A History of Digital Forensics." *IFIP Advances in Information and Communication Technology* 337 (2010): 3–15. doi:10.1007/978-3-642-15506-2_1. <http://www.springerlink.com/content/gr3v34n5248r7x28/>.
- Reith, Mark, Clint Carr, and Gregg Gunsch. "An Examination of Digital Forensic Models." *International Journal of Digital Evidence* 1, no. 3 (2002). <http://www.worldcat.org/wcpa/oclc/223384589?page=frame&url=http%3A%2F%2Fwww.ijde.org%2F%26checksum%3D85756f448e9f3f33b58f16d99aa26bcf&title=&linktype=digitalObject&detail=>.
- Reside, Doug. "Digital Forensics, Textual Criticism, and the Born Digital Musical." Presented in London, England, 2010. <http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-739.html>.
- . "Howard Ashman and Our Digital Future." 2011 <http://www.nypl.org/blog/2011/04/07/howard-ashman-and-our-digital-future>.
- Rogers, Marcus K., J. Goldman, R. Mislán, T. Wedge, and S. Dabrota. "Computer Forensic Field Triage Process Model." In *Proceedings of the Conference on Digital Forensics, Security and Law*, 2006, 27–40.
- Sammes, A. J., and Brian Jenkinson. *Forensic Computing*, 2nd ed. Springer, 2007.
- Scanlan, Daniel M. *Digital evidence in criminal law*. Aurora, Ont.: Canada Law Book, 2011.

- Schofield, D. "Graphical Evidence: Forensic Examinations and Virtual Reconstructions." *Australian Journal of Forensic Sciences* 41 (2009): 131–145.
- Sommer, Peter. "Computer Forensics: An Introduction." In *Compsec Proceedings 1992*. Elsevier, 1992. <http://www.bookdepository.co.uk/Compsec-Proceedings-1992-Monk/9781856171687>.
- . "Digital Footprints: Assessing Computer Evidence." *Criminal Law Review* (Special Edition) (1998): 61–78. <http://www.pmsommer.net/page7.html>.
- Spenneberg, Ralf. "Undeleted: Carving Tools Help You Recover Deleted Files." *Linux Magazine* (93) (August 2008): 30–33.
- Stanley, Philip M. "Computer Crime Investigation and Investigators." *Computers & Security* 5 (1986): 309–313. <http://www.sciencedirect.com/science/journal/01674048>.
- Taylor, Carol, Barbara Endicott-Popovsky, and Deborah A. Frincke. "Specifying Digital Forensics: A Forensics Policy Approach." *Digital Investigation* 4(Supplement 1) (2007): 101–104. <http://www.sciencedirect.com/science/article/B7CW4-4NYD8T4-2/2/42ca628dc0b4d15097058816a107e2b9>.
- US Department of Justice. "Electronic Crime Scene Investigation: A Guide for First Responders." NIJ Special Report. Washington, DC, 2001. www.ojp.usdoj.gov/nij.
- Whitcomb, Carrie Morgan. "An Historical Perspective of Digital Evidence: A Forensic Scientist's View." *International Journal of Digital Evidence* 1, no. 1 (2002). <http://www.utica.edu/academic/institutes/ecii/publications/articles/9C4E695B-0B78-1059-3432402909E27BB4.pdf>.
- Wilding, Edward. *Computer Evidence: a Forensic Investigations Handbook*. Sweet & Maxwell, 1996.

Giving a Permanent Digital Voice to the Silenced

For the Children Taken ...

The Challenge to Truth Commissions in Building digital collections for research and long-term preservation

Terry Reilly¹

*Manager, Document Acquisition and Collections, Truth and Reconciliation Commission of Canada,
University of Calgary, Canada*

Abstract

This paper presents a case study of efforts to fulfill the Mandate of the Indian Residential Schools Truth and Reconciliation Commission (TRC) which includes among its primary goals the responsibility to e) Identify sources and create as complete an historical record as possible of the IRS system and legacy. The record shall be preserved and made accessible to the public for future study and use. This paper explores the challenges in building collaborative teams across corporate and government cultures, and the results of negotiating the concept of “relevance” across differing private and public archival cultures, and ensuring authenticity and discoverability when we present dynamic video within archival descriptive systems. The paper will include an interactive demonstration of the results achieved in our on-line OLTRC databases. In order to fulfill a requirement to create a permanent repository the Commissioners have initiated a process to establish a National Research Centre and ensure the preservation of its archives. The paper presents some of the challenges of moving the collections into an as yet unspecified archival environment.

Author

Terry Reilly has been Manager of Document Acquisition and Collections for the Truth and Reconciliation Commission of Canada since December 2010. She serves the Commission on Interchange from her academic position as an archivist at the University of Calgary, Alberta, Canada. She has served as Director of Archives and Special Collections at the University of Calgary and as General Synod Archivist of the Anglican Church of Canada. She is a past president of the Association of Canadian Archivists and currently serves as the chair of the Archivists of Religious Collection Section of the Society of American Archivists. She holds an M.A. in history from York University in Toronto and has been a Bentley Fellow.

“The road we travel is equal in importance to the destination we seek. There are no shortcuts. When it comes to truth and reconciliation, we are all forced to go the distance.”

-Justice Murray Sinclair,

*Chair of the Truth and Reconciliation Commission of Canada, to the Canadian Senate
Standing Committee on Aboriginal Peoples, September 28, 2010*

1. Introduction

While most Truth Commissions are established by governments in a transition from authoritarian to more democratic rule² the Canadian Truth and Reconciliation Commission has been established as part of the

¹ Any opinions and all errors herein are my own and not those of the TRC.

² See Trudy Huscamp Peterson, *Final Acts: A Guide to Preserving Records of truth Commissions* (Washington, D.C. and Baltimore, MD: Woodrow Wilson Centre and Johns Hopkins University Press, 2005).

Settlement agreement of the largest class action law suit in Canadian history. Although has been no regime change in Canada, the Commission is similar to others in that its mandate is time limited and the three Commissioners are required to produce a final report.³

2. Background⁴

Up until the 1990s, the Canadian government, in partnership with a number of Christian churches, operated a residential school system for Aboriginal children. These government-funded, usually church-run schools and residences were set up to assimilate Aboriginal people forcibly into the Canadian mainstream by eliminating parental and community involvement in the intellectual, cultural, and spiritual development of Aboriginal children. More than 150,000 First Nations, Inuit, and Métis children were placed in what were known as Indian residential schools. As a matter of policy, the children commonly were forbidden to speak their own language or engage in their own cultural and spiritual practices. Generations of children were traumatized by the experience.

The lack of parental and family involvement in the upbringing of their own children also denied those same children the ability to develop parenting skills. There are an estimated 80,000 former students still living today. Because residential schools operated for well more than a century, their impact has been transmitted from grandparents to parents to children. This legacy from one generation to the next has contributed to social problems, poor health, and low educational success rates in Aboriginal communities today. The 1996 Canadian Royal Commission on Aboriginal Peoples⁵ and various other reports and inquiries have documented the emotional, physical, and sexual abuse that many children experienced during their school years. Beginning in the mid-1990s, thousands of former students took legal action against the churches that ran the schools and the federal government that funded them. These civil lawsuits sought compensation for the injuries that individuals had sustained, and for loss of language and culture. They were the basis of several large class-action suits that were resolved in 2007 with the implementation of the Indian Residential Schools Settlement Agreement, the largest class-action settlement in Canadian history. The Agreement, which is being implemented under court supervision, is intended to begin repairing the harm caused by the residential school system.⁶

In addition to providing compensation to former students, the Agreement established the Truth and Reconciliation Commission of Canada with a budget of \$60-million and a five-year term. The Commission's overarching purposes are to: reveal to Canadians the complex truth about the history and the ongoing legacy of the church-run residential schools, in a manner that fully documents the individual and collective harms perpetrated against Aboriginal peoples, and honours the resiliency and courage of former students, their families, and communities; and guide and inspire a process of truth and healing, leading toward reconciliation within Aboriginal families, and between Aboriginal peoples and non-Aboriginal communities, churches, governments, and Canadians generally. It is hoped that the process will work to renew relationships on a basis of inclusion, mutual understanding, and respect.

³ The Mandate of the Truth and Reconciliation Commission is found in *Schedule N of the Indian Residential Schools Settlement Agreement* see www.trc.ca/OurMandate.

⁴ *Truth and Reconciliation Commission Interim Report*, Truth and Reconciliation Commission of Canada, Winnipeg and Ottawa, 2012. This publication and its related history book *They Came for the Children: Canada Aboriginal Peoples and Residential Schools*, are freely available at www.trc.ca.

⁵ See *Report of the Royal Commission on Aboriginal Peoples* at <http://www.aadnc-aandc.gc.ca/eng/1307458586498>.

⁶ See <http://www.residentialschoolsettlement.ca/english.html>.

While the residential school system operated across Canada, the majority of schools were located in the West and the North. For this reason, the Commission established its head office in Winnipeg, Manitoba. It retained a small Ottawa office, and opened satellite offices in Hobbema, Alberta, to support Commissioner Wilton Littlechild and Yellowknife, Northwest Territories to support Commissioner Marie Wilson and the Inuit Sub-Commission. Chairman and Commissioner Mr. Justice Murray Sinclair's office is in Winnipeg. To extend the Commission's reach into smaller centres and communities and as required by the Settlement Agreement, regional coordinators work to support local Statement Gatherers and related Hearings and also relate to independently organized Community Events.

In recognition of the unique cultures of the Inuit, and the experiences and impacts of residential schools on them, the Truth and Reconciliation Commission also established an Inuit Sub-Commission. It is charged with ensuring that the Commission addresses the challenges to statement gathering and record collection in remote, isolated Inuit communities, and among Inuit throughout Canada. The Inuit Sub-Commission provides the environment and supports necessary to earn the trust of Inuit survivors.

The Commission staff is drawn from the public service, private sector, and non-governmental organizations. As of August 2012, the Commission employed 55 people and 60% are aboriginal employees who work at all levels of the organization. There are three archivists and one part-time video editor in the Document Collections section of the Statement Gathering area.

The Mandate of the Indian Residential Schools Truth and Reconciliation Commission (TRC) has among its primary goals the responsibility to *Identify sources and create as complete an historical record as possible of the IRS system and legacy. The record shall be preserved and made accessible to the public for future study and use.* The work of the Archivists in the Commission's Document Collection team is directed towards fulfilling this Mandate.

3. "Telling Your Truth" - the Commission's work of Statement Gathering

Until now, the voices of those who were directly involved in the day-to-day life of the schools, particularly the former students, have been largely missing from the historical record. The Commission is committed to providing every former residential student—and every person whose life is affected by the residential school system—with the opportunity to create a record of that experience.

The work of other truth and reconciliation commissions has confirmed the particular importance of the statement-giving process as a means to restore dignity and identity to those who have suffered grievous harms. Statement gathering is a central element in the Commission mandate, and statement giving is voluntary with complicated logistics. The statements gathered are being used by the Commission in the preparation of its reports, and eventually will be housed in the National Research Centre. Statement gathering has occurred at National Events, Hearings and Community Events.

The Commission has provided for three distinct statement gathering opportunities/venues. Private Statements are given by one individual to another. Sharing Circles are done in groups with a moderator. They can be as large as 10 or 12 people. Circles are public statements while anyone can witness or listen although they are sometimes conducted without an audience. Sharing Panels are statements given in public places at public events to the TRC Commissioners. All Sharing Panel statements are public and anyone can listen/watch. Members of the Media may also be present at the panels.⁷

⁷ Definitions supplied by Raegan Swanson, TRC archivist, 23 August, 2012.

4. The Statement Gathering Processes – Archival at the Beginning

Statement providers are encouraged to talk about any and all aspects of their lives they feel are important, including times before, during, and after attending a residential school. The family members of survivors, former staff, and others affected by the residential schools also are encouraged to share their experiences.

The Commission recognizes that providing a statement to the Commission is often very emotional and extremely difficult. For this reason, statement providers are given the option of having a health support worker, a cultural support worker, or a professional therapist attend their session. These health supports ensure statement providers are able to talk to someone who can assist them if necessary before and after providing a statement. Individuals are given the option of having an audio or video recording made of their experiences. If they wish, they are given a copy of their statement immediately at the end of the interview. They may choose to provide their statement in writing or over the phone if proper health supports are in place. All individual statement providers must choose whether or not their statement will be retained as confidential or made available for research.

At Sharing Circles and Commission hearings, statements are made in a public setting. People who make their statement in a private setting can choose from two levels of privacy protection. The first option ensures full privacy according to the standards of the federal *Privacy Act*. The second option allows the statement provider to waive certain rights to privacy in the interests of having their experiences known to, and shared with, the greater public.

People who waive those rights are giving consent to the Commission and to the National Research Centre to use their statement for public education purposes or to disclose their statement to third parties for public education purposes in a respectful and dignified manner (such as for third-party documentary films). The Commission and National Research Centre have the authority to decide whether to provide such access. These options are explained carefully to the statement provider before a private statement-gathering session. To date, over half the statement providers have chosen to have their statements recorded for public education purposes. The Commission also ensures that all digital information is transmitted and protected carefully during trips in and out of the field.⁸

The Commission has made it a high priority to gather statements from the elderly or ill, as well as from particularly vulnerable and marginalized former students who are at risk. It has undertaken a number of innovative measures, including a day-long event facilitated by Métis Calgary Family Services at the downtown branch of the Calgary Public Library that focused on collecting statements from homeless individuals. Projects designed to reach those survivors in jails also are underway.⁹

By the end of August 2012, the Commission had collected 2102 private statements.¹⁰ An additional 1374 statements had been given in Sharing Circles and at public hearings. One hundred and fifteen material and artistic submissions had been received. The Commission now has both the mechanisms and process in place to ensure it is able to meet its statement-gathering goals. Regional liaisons play a role in coordinating and organizing a series of specific and targeted visits to communities across the country.

⁸ The Statement Gathering process is governed by The TRC's Statement Gathering Manual and related Supporting Document forms that are developed and controlled by the Archivists. All this documentation is available on the TRC's shared server. Archivists hold regular in person and on-line training sessions and are available for consultations.

⁹ http://www.wawataynews.ca/archive/all/2012/8/16/residential-school-impacts-still-seen-kenora-jail_23298.

¹⁰ Statistics are gathered on a monthly basis for planning purposes.

5. The Settlement Agreement and the Churches who managed the schools

The group of Christian entities known collectively as the historic mission Churches (Anglican, Presbyterian, United and Roman Catholic—including religious orders and dioceses) as signatories to the Settlement Agreement are required to provide the Commission with “all relevant documents relating to the history and or legacy of Indian Residential Schools in Canada.” The first project related to this work was the establishment of an Archivists Working Group. The Group was concerned about identify the issues of concern regarding the portion of documents where there were perceived competing legal obligations; identifying how many documents there were in this category; and, determining solutions to deal with the issues of the implied undertaking in the copies Church records located in the federal NRA collection; questions related to solicitor-client privilege in both federal and Church documents; the application of provincial privacy and Charter law to individuals named in the process; and privacy law more generally.

In 2009-2010 in order to support these discussions and to determine the scope of the project a survey of Church Archives was conducted. Also, the Parties to the Settlement Agreement discussed but did not come to total agreement about the definition of relevance.¹¹ In December 2010 as Manager of Collections I decided that I would proceed with a working definition provided by the archivist of the Presbyterian Church in Canada. The Commission seeks to acquire a comprehensive collection of all relevant documents in any format related to the history and or legacy of Indian Residential Schools in Canada. While this definition does not solve the problem it has allowed the work to proceed.

6. The RFP for the Digital Project

In the fall of 2010 the Commission posted a Request for Proposals on the Government of Canada’s procurement site. The successful applicant was to provide all required , project design and management, human resources, computer hardware, software digitization and meta data expertise to create a digital repository of data bases (initially referred to as a data base) of historical records related to the Indian Residential Schools system. The repository was to be designed include both Government of Canada and Church records. The project team was required to have the capacity to locate records across the country. This was based on information provided by the Church Entities and in particular volume estimates stated in the preliminary surveys that had been submitted to the Archives Working Group. These surveys also showed there were a variety of locations both large and professionally staffed to very small archives, to assist both professional and volunteer whether full-time or more often part-time archivists with their task of identifying records, to describe these records for the Commission, to create digital images and their associated meta data which would enable research by the TRC’s research team and any other researchers contracted by the Commission. The results of the project had to be designed so that they could be made accessible (to the extent that privacy and other related law allows) for public research at the mandated National Research Centre. Bronson Consulting Group of Ottawa and its partners, The History Group, Brechin Imaging and Minisis, Inc., were the successful bidders. There are several parts to the project which has evolved from a single data base platform to a digital repository as the project has unfolded over the past 18 months.

¹¹ The All Parties did determine that “Relevant information is information that is necessary to fulfill any of the mandated goals and activities of the TRC and provided a list of records it felt would satisfy that definition. See Tom McMahon, “Records of Discussion with the All Parties Group,” January–February, 2010.

The Church Archives Project is designed to collect records in either digital or photocopy form and works with six major functions: Contact with the Archives; File Identification; File Review; Digitization; Postprocessing and Metadata tagging. Each of these steps poses its own challenges related to time and budget.¹²

While in general Church archivists have been supportive many are part-time volunteers and some signatories to the Agreement have no personnel with archival training. Many archives are only open for a few hours per week, though in most cases archivists have accommodated the TRC contractors hours. To ensure the best success in year one the TRC deliberately worked with the largest archives and in particular with those that had “national responsibilities.” In addition regional or local archives in British Columbia both Anglican and United Church and the Jesuits of Upper Canada served as pilot sites for smaller archives. File identification is based on finding aid review. Challenges have included incomplete or non-existent finding aids; differing versions between paper and electronic finding aids and in particular paper often hand written finding aids which were known to be out of date. The biggest challenge both for the archivists and the TRC contractors identifying relevant documents are large collections of unaccessioned or unorganized records. Also we did find that in general the Commission’s choices were broader than the archivists though in most cases the archivist did agree to digitization some records have been held back. In the case of records held back under solicitor-client privilege the signatory will provide a list to the Commission’s legal counsel. In order to ensure the highest standards were consistently maintained at every site a detailed set of protocols was developed for on-site work, document handling and finding aid recording. In addition digitization equipment was chosen to ensure that the specific needs of each archives were respected. Explicit folder and documents identification naming conventions were applied to all documents removed from files for digitization. While the option for sheet feeding was available the Church archivists were very reluctant to allow the Commission’s contractors use it even for records that were produced from digital files with the exception of photocopies that had previously been used in litigation. After digitization post processing ensured that the archival relationships between documents were consistently maintained but this did add significant costs and slowed the process of uploading collections into the database. Although significant effort had been made to develop a extensive set of key words based on the Commission’s research requirements it turned out that financial constraints forced the decision to metadata tag only a test subset of the documents from one site and one large set of photographs that had no finding aid. In order to overcome this handicap the Director of Statement Gathering has overseen the development of an on-line search system that is available for the Research Directorate. At the end of the pilot phase there are 202,787 pages of Church records from nine sites, and 980,000 Government of Canada documents received from the NRA system including 129 School Narratives.

The second year of the project has been much more modest. Financial constraints have meant that a new program was developed that limits the Commission’s work to finding aid identification and ingest and returns the work of file review and document production to the Church Archives. The Commission is able to receive either digital or photocopied collections. At this time only digital collections will be added to the repository. Photocopied collections are being accessioned and will be held in the Commissions secure storage.

¹² Martin McGarry, Program Manager Bronson Consulting Group, “TRC Database Project Overview,” TRC Senior Executive Briefing, 4 January 2012.

7. The TRC's digital repository – MINISIS

For now access to the TRCs digital repository is limited to the Commissioners authorized researchers and documents collection staff. At present the repository consists of separate but linked databases for the Church Records, the National Research and Analysis hereafter NRA data transmitted from Aboriginal and Northern Affairs Canada, the Schools data base which contains base line bibliographic and historical data, and the data base for Statements known as VIMS.

The TRC data base partner for this project is Minisis, Inc. There are 2 major sets of functionality in the system: Project management functionality and Data Collection functionality (specifically file identification, meta data tagging and searching). Project management functions include, the tracking and authorization of jobs, workflow which links Activity assignments to the Data Collection functions, estimating and costs and operations controls. Data Collection functions ensure that data entry occurs in an organized fashion and that records can be accessed hierarchically via any of the data sets.

8. Uploading private statement access copies to the digital repository VIMS

Most private and public statements that the TRC has collected were recorded in HD video or in Broadcast Wave format for audio, resulting in very large master media files (for example, 12 minutes of HD video is roughly 2GB in size). These master files are too large to be streamed online from VIMS or easily downloaded. It is therefore necessary to create smaller access copies in compressed audio and video formats (MP3 format at 128kbps for audio files, and MP4 format with H.264 video codec and MP3 audio for video files). These access copies (also called derivatives) are usually about 5% the size of the original master files.

Once created, these access copies are sent to Minisis for upload to VIMS. Like the transcripts, all access copies are given standardized filenames based on the private statement filename (for example, the video access copies associated with private statement 2011-0123 would be titled 2011-0123.1.mp4, 2011-0123.2.mp4, etc.). This allows Minisis to run a script that automatically uploads the access copy media files and link them directly to the VIMS entry of the private statement they are associated with.

9. Creating transcripts of Private Statements and uploading them into the VIMS repository

In the summer of 2012 to assist TRC authorized researchers who require transcripts of Private Statements, we designed an in-house project and contracted with a team of four to create transcriptions. In order to ensure consistent practice the archivists wrote a transcription manual for the contractors to use.

Only non-protected, i.e., Private Statements that are fully available for direct research and quotation in the Commissions publications are being selected for transcription. The archives staff do not have the resources to redact or anonymize statements that have been restricted by the Statement giver. While the selection of private statements for transcription is to be determined by the current needs of the TRC's Researchers the Archivists also select statements for transcription that an approximately even number of statements from each geographical region will be transcribed. Transcribers use a pair of headphones, a Philips Foot Control LFH 2330/00 foot pedal, and a computer work station with Express Scribe Pro software and Microsoft Word installed. Transcribers follow procedures and rules set out in the manual

and are laid out in the transcription template. Once completed and reviewed by one of the archivists transcripts are saved on the Archives server in the same computer folder as the associated private statement's media files and supporting documents. An access copy of all private statement transcripts is also be saved on the TRC's shared drive. The TRC's Researchers are notified when new transcripts are posted to the shared drive.

For purposes of ingest into the VIMS repository 2 copies are made, one a PDF and the other a Word document. These copies are sent to Minisis for upload to VIMS. The text of the copy in Word format is directly copied to a data field in the VIMS entry for the associated private statement; this allows for keyword searching of the entire transcript via the VIMS search interface. Copying the transcript text to a VIMS data field removes all rich text formatting, so it is also necessary to upload and attach the PDF version of the transcript to the VIMS entry. This PDF can be viewed or downloaded by the research team with all the original text formatting intact. We chose PDF as the access copy format due to its widespread use and the security the format offers against unauthorized editing.

Uploading the Word and PDF versions of the transcripts to VIMS is an entirely automated process handled by Minisis, as their programmers have created scripts to extract the text from the Word document and paste it into the appropriate data field in VIMS, and also to upload and link the PDF version of the private statement to its description in VIMS. All that is necessary for this automated process is for me to provide all transcripts with a standardized filename based on the unique statement number. For example, as long as the two versions of the transcript for private statement 2011-0123 are titled 2011-0123.doc and 2011-0123.pdf, they will be automatically processed and uploaded to VIMS by Minisis' scripts.

10. The Settlement Agreement and the Government of Canada – the National Research and Analysis (AANDC) processes

As of July 2012 the major addition to the TRC's digital repository is the NRA (National Research and Analysis directorate of Aboriginal and Northern Affairs Canada, formerly the department of Indian Affairs) database. This database contains historical and contextual records that Canada collected that were relevant to the numerous residential schools litigations. The research and document collection processes differed substantially among litigation offices in eastern Canada, Alberta and western Canada which has resulted in an uneven collection. The Canadian Judicial Council Standard for electronic document transmission in litigation was developed over a number of years and is now the accepted metadata standard for projects of this kind. Provenance or collection level meta data is often not longer visible in the meta data though it may be visible on the individual digital documents. Just over a thousand items have been identified as having meta data but no associated image.

This database contains approximately 1.2 million records, and is made up of both Canada's (including AANDC records held at LAC) and church records. While the records in this database do not contain a complete record of the residential school system, the database is of enormous value to the TRC for both operational and historical purposes.

Also Library and Archives Canada have provided copies of their Red, Black and Schools series for the Repository. These records are currently being analysed and a process for ingesting them into the Minisis system is being developed.

11. Still to Come

Under the general direction of staff of Aboriginal Affairs and Northern Development Canada (the Department of Indian Affairs) surveys are being completed of semi-active and inactive records that relate to residential schools and its legacy but that either are not scheduled or have not been transferred to Library and Archives Canada. The Commission is expecting to receive 33 collections from this program. Discussions with the federal government are also continuing with regard to the records of 2 other Residential Schools programs, the Common Experience Payments (CEP) and the Independent Assessment Process (IAP). Finally the Commission is conducting research into the extent to which records held by Library and Archives Canada are not currently represented in the Collection.

12. Challenges for the National Research Centre (NRC)

To assist the archivists in the future NRC the TRC archivists are compiling a detailed policy and procedures manual in addition to the Statement Gathering Manual and the Protocols and technical reports that the Bronson Consulting Group have developed and submitted. In addition all of the current research into trusted digital repositories will be essential to ensuring that embedded technical specifications can be used to support data migration. A robust privacy regime with specialist assistance will be essential to ensuring that the rights and responsibilities of the donors including statement providers and the contributing archives and government departments are maintained. It will be essential to ensure appropriate culture respect during research special rooms for ceremony and reflection. The archives of the Museum of the American Indian provides a model.¹³ Finally, the logistical challenges that the TRC has faced will continue as the Centre works with local and regional communities to provide educational resources and access to this unique memorial collection.

¹³ See <http://nmai.si.edu/explore/collections/archive/>.

National Strategies as the Foundation of Togetherness

National Planning as the Key for Successful Implementation of Digitization Strategies

Andris Vilks and Uldis Zariņš

The National Library of Latvia, andris.vilks@lnb.lv

Abstract

Digitising Latvian cultural heritage and making it publicly available is one the most important tasks for all Latvian cultural and memory institutions. Unfortunately, there are some legislation problems that cause lack of coordination and joint methodology in the digitization field. The National Library of Latvia has taken part in development of a state “Digital cultural heritage” working plan that foresees establishing of the digitization competence centre. Effectiveness of cooperation and sharing of best practices was approved by experience of many European countries. It helps not only to reduce costs and save resources, but also facilitates development of all cooperation partners.

Authors

Andris Vilks is one of the most prominent specialists of the library and information sciences. He has contributed significantly to the development of the library sector and advancement of information and knowledge society in Latvia and the Baltic States. Born in 1957 in Riga, graduated the Doctoral program of the Library and Information Sciences at the University of Latvia. Since year 1978 he works in the National Library of Latvia (NLL) and since year 1989 he is the Director of the National Library of Latvia. Since 1989 Andris Vilks has ardently and tirelessly advanced the project of the National Library of Latvia “The Castle of Light”, which includes the construction of the new building of the library, development of the country wide library network and establishment of the National Digital Library. Mr. Vilks holds different positions in national and international organizations. Andris Vilks is Chancellor of Latvian State Decorations Chapter, the Treasurer and the Member of Executive Committee of the Conference of European National Librarians (CENL), the Member of International Committee of Experts of UNESCO Memory of the World Register (1992 – 2007, 2009 – present), the Member of the Executive Committee of Latvian National Commission for UNESCO, President of Latvian National Commission for UNESCO (2008 – 2010) and since 2001 Chairperson of Latvian National Commission for UNESCO Communication and Information Programme.

Uldis Zariņš is Head of Strategic Development.

*“Member States should improve how they preserve digital material for the long-term.
Otherwise our collective memory will be lost for future generations.”*

-- Neelie Kroes

Vice-President of the European Commission responsible for the Digital Agenda

Many institutions involved in national cultural heritage digitization have realised that they do not have enough financial and human resources to succeed, so they have started to cooperate with other institutions.

Indeed, cooperation on preservation infrastructures, use of open-source software and sharing of best practices have proved very efficient for cost and resource saving, as well as for helping organizations to move forward more swiftly. Cooperation has also flourished on an international level. Such successful initiatives are “Europeana” and “World Digital Library” (WDL). However, there is a cooperation level

where surprisingly often is not enough collaboration—the national level. Of course there are some exceptions. An excellent example of collaboration on a national scale is the National Digital Library of Finland, when libraries, archives and museums have joined their resources to achieve common goals. In spite of that, more common practice is inter-branch cooperation.

Digitising Latvian cultural heritage and making it publicly available in the digital environment currently is one the most important tasks for all Latvian cultural and memory institutions. This objective helps to update cultural institutions' workflows and offer quality services to target groups, which primary use digital media for information search and retrieval, as well as for communication processes. At the same time such improvements ensure cultural heritage long-term preservation for future generations.

The importance of digitising Latvian cultural heritage and making it publicly available in the digital environment is mentioned in the Republic of Latvia long-term strategic planning documents ("National Culture Policy Guidelines 2006-2015. National State" and "Latvian National Development Plan 2007-2013," as well as outlined in the 2014-2020 Plan). In 2006 for the implementation of the digitization assignments a long-term government programme "Latvian National Digital Library "Letonica"" was started. Starting with the 2006 for the development of "Letonica" financial resources from the state budget sub-programme "National United Library Information System" are allocated. The most significant accomplishment of this programme is the establishment of the Digital Object Management System (DOM) that was started in 2007 and finished in 2009. At the same time when DOM technical infrastructure was being built, within the framework of this programme, digitization of the National Library of Latvia (NLL) collections was done and digitization methodology base was developed. Since 2009 NLL has found the opportunity to continue the digital library development process by implementing two European Regional Development Fund financed projects and a chain of small-scale projects sponsored by foreign financial instruments.

At the beginning of the programme implementation process, it was planned that project outcomes (technological and methodological base) will be used not only by NLL, but also by other libraries, archives and museums. In reality, apart from the advanced NLL digital library platform many information systems that are partly connected with completing digitization tasks were developed. For example, digital object storage and display functionality of the "Joint Catalogue of the National Holdings of Museums" and "State Unified Information System for Archives," digital object search functionality of the "Cultural Portal," purchase of digitized data storage and management system for archives and museums, development of digitization methodology for archives, etc. These products were developed without taking into account what NLL had accomplished during the National Digital Library development process. As a result of memory institutions stand-alone activities, incompatible and even duplicate information systems were established. Implemented technical solutions and projects of archives and museums are at odds with not only NLL digitization methodology, but also international standards.

The main reason for such lack of uniformity is that no institution in Latvia has a formal mandate that delegates the functions of creating and maintaining a national scale digital library. As a result of legislation problems, digitization activities are being done within certain institutions and domains, without coordination, broader view of governmental digitization objectives, joint methodology and standards, and without noticing achievements of related institutions. The NLL has considered such use of state and European Union (EU) resources unsuitable, and suggested that the Ministry of Culture start digitization coordination process.

Further development of digital process coordination on a national level is extremely important, because publicly available information indicates that digitization and digital availability will also be

among the European Commission (EC) priorities for the next financial planning period. The European Digital Library “Europeana” sustainable development is one of the aims of the EC programme “Digital Agenda for Europe.” EC report “Europeana – next steps” and EC established “Comité des Sages” (Reflection Group) report “The New Renaissance” emphasizes that EC must continue financing the maintenance of “Europeana”, while the creation of the digital content expected to be done by member states, especially encouraging to realise digitization projects with public-private partnership support or using the structural funds resources.

The NLL has decided that the most effective way to achieve maximum rationality and effective use of resources is by initiating National digital resource aggregator development, which would be based on the previous projects outcomes, first of all, on the “Latvian National Digital Library “Letonica”” programme. This initiative could unite competences and resources of the leading Latvian cultural and memory institutions. The NLL is ready to undertake the task of establishing and coordinating the aggregator development, because NLL has the most significant experience and knowledge in the field of digitization in Latvia. NLL has started creating the concept of digital content aggregator that expected to be completed by the end of 2012.

NLL has been creating digital resources since 2001, when it also initiated cooperation on digitization between libraries, archives and museums. The main functions of NLL are: responsibility for information gathering, processing, storage and availability, as well as hard work on maintaining and developing technologies and human resource competencies, which allows NLL to take responsibility of establishing the united concept for aggregator functionality, developing digitization methodology, implementation of international standards, personnel training and cooperation with international organizations, including the European Digital Library “Europeana”. It should be noted that NLL created “Digitaser’s Manual” is a major digitization teaching aid in Latvia. NLL is the only institutional content aggregator for “Europeana” and WDL in Latvia. NLL took part in many successful “Europeana” digital content development projects, in close cooperation with different international organizations: Conference of European National Librarians (CENL), Conference of Directors of National Libraries (CDNL), Ligue des Bibliothèques Européennes de Recherche—Association of European Research Libraries (LIBER). The NLL staff regularly takes part in “Europeana” Council of Content Providers and Aggregators working groups, International Federation of Library Associations and Institutions (IFLA) and European Bureau of Library, Information and Documentation Associations (EBLIDA) copyright committees. The NLL professionals actively support the Ministry of Culture of the Republic of Latvia, help to prepare national position on the issues related to digitization of cultural heritage and adjust the Republic of Latvia legislation on related issues—from the legal deposit to copyrights.

At the same time NLL is ready to closely cooperate with other institutions which competencies at some areas are stronger than NLL’s ones. One of the most important digitization aspects is long-term digital content preservation that traditionally is the archives’ competence. It could not be denied that state agency “Culture Information Systems” has significant experience in building and maintaining IT infrastructure and information systems. This valuable experience could be used to maintain also a digital library infrastructure.

The NLL emphasizes that this initiative, as well as programme “Latvian National Digital Library “Letonica””, does not just mean NLL collection and electronic service development, but step-by-step memory institutions resource integration and maximum effective use of different institutions’ resources and competencies. It is the reason, why NLL have considered that it will not be right to use budget sub-programme “National United Library Information System” for digitization, building and maintenance of

digital library infrastructure. The NLL has advised to establish a new budget programme “Digital Library” that could provide financial resources not only for developing information systems, building and maintaining infrastructure, but also for providing digital services and creating digital content. Referring to positive experience from 2003 until 2005, when “Cooperation of Archives, Museums and Libraries in Digital Environment” was functioning, NLL recommends to renew this practice.

Effectiveness of nationally coordinated digitization processes was approved by experience of many European countries, especially by countries with relatively low inhabitant number. One of the best examples is Finland, where national aggregator development coordination was delegated to the National Library of Finland, but preservation tasks to the National Archives of Finland. The Ministry of Culture of the Republic of Estonia in the end of summer 2012 approved digitization development plan “The National Library of Estonia Digital Archive: Development Plan for 2011-2016.” It was figured out that the digitization processes in Estonia had not been well coordinated so far. The National Library of Estonia initiates the development of unified long-term digital storage, joint Estonian cultural heritage space, as well as strengthening mutual coordination between institutions and establishment of competence centres. 12-13 million euro were allocated to the fulfillment of these tasks during the aforementioned period of time.

The NLL has advised the Ministry of Culture of the Republic of Latvia to consider the proposal of aggregator concept and to start its implementation in 2013, and entrust concept implementation preparatory work, as well as digitization processes coordination, to NLL.

Following objectives are assigned to ensure digitization of all Latvian memory institutions cultural heritage collections:

- Developing of memory institutions collection digitization plans, considering common content selection parameters, where priority is given to the most important collections (perishing, the most demanded and rare – historically significant);
- Establishing of digitization council and competence centre network;
- Developing digitization standards and quality requirements that meet international standards;
- Electronically cataloguing information of memory institutions collections and making it publicly available (according to digitization plan);
- Digitising cultural heritage collections and making them publicly available (including films, sound and audio records, music scores and other materials, according to digitization plan);
- Developing electronic book and periodicals deposit (storage) in cooperation with publishers to ensure long-term preservation of this content;
- Digital recording of intangible cultural heritage content (performances, concerts, oral history etc., according to digitization plan).

The following objectives are assigned to ensure digitally born culture heritage acquisition and storage:

- Developing guidelines and system for acquisition and storage of digitally born and unpublished traditional form materials (e-books, e-publications, digital photos, pictures, electronic maps, video and audio recordings);
- Developing guidelines and system for acquisition and storage of wide variety non-commercial software products (different applications, games etc.);

- Developing guidelines and system for acquisition and storage of private collections materials (digitized / digitally born documents, photos, pictures, e-mail correspondence etc.);
- Developing guidelines for harvesting and storage of Latvian online resources, harvesting and preservation of Latvian web content;
- Establishing storage for digital music recordings and scores.

One of the most important objectives is digital and digitally born cultural heritage long-term preservation in the digital environment. The following activities are assigned to ensure this objective:

- Developing a plan for cultural heritage long-term preservation in the digital environment;
- Establishment of a long-term digital preservation competence centre;
- Developing digital and digitally born cultural heritage long-term preservation infrastructure and services.

Digital strategy could include the role of memory institutions as implementers of programmes such as “Language Block” (CLARIN) and linguistic intelligence technologies “Language Coast”. The following activities are done to unite participants of the language project (memory institutions):

- Automatic translation (machine translation technologies, linking up language syntax and semantics, modeling algorithms and newest statistical methods);
- Terminology management (national and international institutions terminology database integration in unified network, multilingual terminology integration to the translation environments and systems);
- Intelligent search systems (information selection according its meaning, picture analysis, search and automatic annotation technologies, video material analysis technologies, information retrieval from multilingual resources);
- Human voice technologies (voice synthesis, speech recognition and transformation into textual information, human voice recognition methodology for adaptation from large to small language groups, identification of the speaker);
- Orthography technologies (word and sentence syntax analysis, error finding and correction technologies);
- Digital library technology development (integration of technologies for processing, archiving and making available resources with huge content amount, finding solution for long-term preservation, developing interactive and multilingual e-training tools, also content development tools for mobile devices);
- Semantic information management (document and content management automatisisation based on semantic analysis).

There are some objectives that strongly require national coordination and planning:

- Planning of digitization on a national level that allows to set goals of national importance, puts digitization in a broader context, involving all cultural sector (not only the memory institutions), ensures the digital availability of the national cultural content and fosters its broad reuse in the education and creative industries;

- Support of cooperation on a national level by establishing content priorities, implementing centralised governance and financing of digitization, by creating competence centres and fostering knowledge transfer, by developing common digitization standards and setting up a shared infrastructure, thus decreasing the digitization costs and improving quality of outcomes. According to unfinished “Digital cultural heritage working plan, 2013” a new function will be delegated to the NLL—serving as a cultural heritage digitization competence centre that coordinates national digitization processes, develops and disseminates best practices, methodological materials, digitization guidelines and standards. This working plan could be completed and approved by the Cabinet of Ministers in autumn 2012. For establishment of digitization centre at least 5 specialists will be required: metadata specialist, text digitization specialist, digital library systems specialist and project manager.
- Better support of cooperation on an international level. For example, identifying content of national interest located abroad and content of interest to other countries located in holdings of national institutions, and agreeing on cooperation in digitization of this content;
- Support by tackling the issues of political and legal nature. For example, by adjusting the copyright framework, legislation on privacy rights and public sector information digitization;
- Planning of adequate financial and human resources for cultural institutions. Too often cultural institutions are obligated to do more and more, including digitization, without increasing available resources. This also includes development of human resource competencies, because digitization requires additional skills from the cultural institutions personnel;
- Increasing awareness of digitization importance. Too often politicians, staff of cultural institutions, creative industries and wider public do not understand why it is beneficial for all to put the culture in the digital environment.

Lately there have been several calls to different states of the world to increase their efforts on digitising their cultural heritage and making this content available online. The “IFLA Manifesto for Digital Libraries,” endorsed by UNESCO, addresses national governments to recognise the strategic importance of digital libraries and asks to actively support their development by providing necessary legislation and financial support. Manifesto recognises national e-strategies as a firm basis for planning digital libraries. EC recommendation and European Council conclusions on digitization and online accessibility of cultural material and digital preservation provide some very clear guidelines for EU member states, highlighting that in order to fully develop the potential of “Europeana” it is necessary to create synergies between national efforts.

European Council asks its member states to consolidate national strategies and targets for the digitization of cultural content, to consolidate organizations and funding of the digitization, to develop qualitative standards for it, to set up or reinforce national aggregators of digital cultural content, to reinforce the use of common digitization standards and the systematic use of permanent identifiers, to ensure the wide and free availability and reuse of existing metadata produced by cultural institutions, to reinforce national strategies for long-term digital preservation, to make all necessary legal arrangements for the legal deposit of digitally born content in order to guarantee its long-term preservation.

These are some very powerful statements, and hopefully governments of the world are finally taking notice.

Preserving the Crowdsourced Memories of a Nation

The Singapore Memory Project

Ivan Chew¹ & Haliza Jailan²

¹Assistant Director, National Library Arts & Singapore Memory Project, National Library Board Singapore

²Haliza Jailani, Assistant Director, Metadata Services, National Library Board Singapore

Abstract

The Singapore Memory Project (SMP) is a national initiative started in 2011 to collect, preserve and provide access to Singapore's knowledge materials, so as to tell the Singapore Story. The paper will share the approaches, considerations and challenges in collecting memory items and multiple agencies, including open submissions from individual members of the public. The paper will also cover how social media engagement for the project has been adapted as part of the collection process—particularly Facebook, Twitter, and Blogs—and the resulting challenges for digital preservation. Finally, the paper presents the on-going efforts in refining the SMP's digital preservation policies and implementation.

Authors

Ivan has been with the National Library Board since 1996. His current portfolio includes the development of web portals, apps and web services to support national initiatives like the Singapore Memory Project. Since discovering blogs and social media in 2004, he has been actively blogging and has helped develop blogs for the National Library Board, Singapore. Ivan has an MSc (Information Studies) from Nanyang Technology University, Singapore, and a 1st Class honours degree from the University of London in Information Systems and Management.

Haliza is responsible for the implementation of metadata for purposes of resource discovery and digital preservation. Previously Programme Manager for NLB's Digital Infrastructure programme which delivered the Digital Preservation System, she had subsequently operationalised the digital preservation processes for NLB's digital collections. She received her MSc from the University of Central England, focusing on metadata studies.

1. Introduction

The Singapore Memory Project (SMP) is a national initiative started in 2011 to collect, preserve and provide access to Singapore's knowledge materials, so as to tell the Singapore Story. It is interested in building a national collection of content in diverse formats (including print, audio and video), to preserve them in digital form, and make them available for discovery and research. The outcome is to be a look-back at Singapore's development, particularly from the views of ordinary individuals, in the form of memories and shared experiences of the nation.

Singapore is a city-state, with a total population of over 5.18 million.¹ It is a relatively young nation, having gained Independence only less than 50 years ago. In recent years, there has been a growing

¹ "Statistics Singapore - Key Annual Indicators," Department of Statistics Singapore, accessed August 27, 2012, <http://www.singstat.gov.sg/stats/keyind.html#keyind>.

awareness on the need to remember where Singapore came from, and how it got to its current stage of development.²

The SMP aims to collect 5 million personal memories as well as a substantial number of published materials on Singapore by 2015. It will do so not only through its own efforts but also via partners and agencies embarking on similar collection drives. It is interested in recollections not just from individuals, but also organisations, associations, companies and groups.

There are also multiple agencies involved, from government, non-government and private-sector organisations. Some of the memory collection drives include schools with the Ministry of Education, working with Non-government organisations like the National Council of Social Service (NCSS) and National Volunteer & Philanthropy Centre. Other government partnerships partner campaigns relating to Defence and National Service, various ministries. Private-sector companies have also been involved, such as banks and home-grown brands.

Overseas Singaporeans and non-citizens are welcome to participate in this project. Citizenship is not a prerequisite for participation. Singapore's memories also include those of tourists, Permanent Residents, students, business people and various travellers who pass by each year.

Since its August 2011 launch, the SMP has collected close to 300,000 personal memories³ and 120 partnerships forged with agencies from the public and private sectors, academic institutions, schools, organisations, clans and communities, as well as niche clubs and interest groups.

2. The Role of the National Library in a Memory Project

The SMP is facilitated by the National Library Board, in partnership with other institutions such as local and overseas libraries, heritage agencies and research institutions. It is led by a steering committee comprising 13 members from academia, media, content owners, community and agencies involved in heritage and national education. These members were chosen to provide a wide representation from various sectors.

The National Library of Singapore, managed by the National Library Board, is the repository of Singapore's publishing heritage. The NLB has garnered positive goodwill and built a credible image over the last decade in transforming the public and national library system in Singapore. It also has extensive partnerships with various government and private-sector organisations.

The genesis of the SMP began as a National Library project to digitize heritage content and materials. It was subsequently positioned and successfully endorsed as a nation-wide memory initiative. A memory project of such a scale would require capabilities in developing skills and systems to collect, process, organise and make publicly available materials. These are traditionally the skill-sets of a library. Under its Library 2010 plan, NLB had put in place an extensive digitisation effort to transit Singapore content from the physical to the digital.

² "Prime Minister Lee Hsien Loong's National Day Rally 2011 (Speech in English), Sunday, 14 August 2011, at University Cultural Centre, National University of Singapore," Prime Minister's Office Singapore, accessed July 14, 2012,

http://www.pmo.gov.sg/content/pmosite/mediacentre/speechesinterviews/primeminister/2011/August/Prime_Minister_Lee_Hsien_Loongs_National_Day_Rally_2011_Speech_in_English.html.

³ Walter Sim, "Memory project draws 300,000 submissions," *The Straits Times*, August 13, 2012, B2 Home.

3. How memories are collected and how content collection has been shaped

The Merriam-Webster dictionary defines “crowdsourcing” as “the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers.” The SMP could be seen as crowdsourced project on a national scale. Both traditional and mainstream channels have their roles as collection channels. Digital and social media channels play an important role too.

The public call for contribution of digital and physical items has been through the following:

- Advertisements/advertorials in print, broadcast and online media
- Social media platforms
- Outdoor publicity, distribution of posters and flyers
- Exhibitions and roadshows at community and residential areas, libraries, schools and partner agencies

The project involves a wide variety of partners and individuals, each with differing skill sets in area of recording, documenting and so on. A large part of the SMP work has been in actively cultivating grassroots communities and volunteers, and to equip people with the skills to conduct interviews and collect memories.

At the social media front, different content posted at the respective social media platforms call for unique approaches and treatment for the acquisition/ collection, organising, storage, preservation, search and display.

Formats collected so far include handwritten notes, personal photographs, physical objects, illustrations, and written expressions in digital and non-digital formats. The presentation, and consequently preservation, is ultimately in a digital form.

The criteria to the types of items that form part of Singapore’s memory have been deliberately broad. The SMP is interested in stories and memories that relate to Singapore, be it past or present memories, in physical or digital format. These may come in the form of photographs, letters, ephemeral, manuscripts, videos, or oral interviews that are stories and memories that are related to Singapore.

For SMP, memories can be personal accounts that tell the Singapore Story, be it past or present memories, in physical or digital formats. The submissions need not just be on the positive or feel-good memories only. The stories could be as recent as something that happened a few minutes ago. What is present now, will serve as a past memory for future generations.

Contributions from partners tend to be publications, many require digitization. For instance, with the Spirit of Singapore NDP engagement programme, more than 200,000 memories have been recorded in the form of reflections of Singapore on the Singapore Spirit. These were articulated entirely as hand-written postcards. They were transcribed, with the postcards scanned as digital images. The transcribed text and the images were then presented as a memory item in the portal.

The collection efforts and drives influence and ultimately have implications for the subsequent processing, organising, retrieval/ display and preservation of the memories. These influence and dictate the quality and quantity, type and form of the content, largely related to the sources, all of which continue to shape the project’s preservation policies and approaches.

4. The SMP digital & social media “eco-system”

The overall approach in SMP has been to reduce barriers to contribution. Where possible, the SMP would go to where users prefer to congregate online, rather than force them to use the portal as the sole submission channel.

The use of social media is seen as a practical reality for online interactions, sharing information, creating and consuming content. A total of 66.6% of Singapore’s resident adults have Secondary or higher qualifications and the country has a literacy rate of 96.1%.⁴ It is also a very connected city, with almost 3 in 4 owning a smartphone⁵ and over 80% of households in Singapore have home access to the Internet.⁶ In that context, it was quite natural that the SMP would have an online and social media component to it.

The current SMP digital and social media “eco-system” comprises of the flagship portal and a blog, as well as official channels on social media platforms like Facebook, Twitter, Instagram, and YouTube (these are a few of the main social media platforms). The social media channels, from their initial roles as engagement platforms, are now positioned as complementary extensions to collect memories online.

4.1 Singaporememory.SG - SMP portal

The SingaporeMemory.SG was launched in March 2012. It was envisioned that every Singaporean would have a personal digital memory account to contribute their stories. With that account, users could connect with others through shared memories.

The SMP portal was developed primarily as a way to efficiently collect crowdsourced content online, and to allow public access to all knowledge assets collected by the SMP from other sources, including institutions. It was also to be a showcase for evocative videos, photo essays, comics and other visual productions from students, artists and filmmakers. Such content was to serve as “emotive hooks” that would make the content alive. It was hoped that users would be invested to discover more, and augment and create further content.

As mentioned, the reduction of barriers to contribution was a constant consideration. The challenge was to build a portal that was on par with popular social media platforms, and still be able to obtain information and formats that would allow the content to be adequately preserved digitally.

One of the ways in which this is done was to let users log in with their existing accounts on popular platforms, like Facebook, Yahoo, and Google. Users could then proceed to contribute to the portal without having to create additional online accounts.

During the design phase, the submission fields were refined to what was considered the minimum required. Compulsory fields were also kept to a bare minimum. It was also decided to provide only one form that would allow different permutations of a memory to be submitted. Examples: Text description-only; Text description with Image/ Video; Image/ Video only.

The current online submission form comprises of the following fields:

⁴ “Statistics Singapore - Key Annual Indicators,” Department of Statistics Singapore, accessed August 27, 2012, <http://www.singstat.gov.sg/stats/keyind.html#litedu>.

⁵ “Singapore Tops Smartphone Ownership,” *AsiaOne*, accessed Jul 14, 2012, <http://news.asiaone.com/News/Latest%2BNews/Science%2Band%2BTech/Story/A1Story20120619-353711.html>.

⁶ “IDA Singapore - Publications - Infocomm Usage - Households and Individuals,” IDA Singapore, accessed August 12, 2012, <http://www.ida.gov.sg/Publications/20070822125451.aspx>.

- Title of the memory (compulsory field)
- Year (compulsory field)
- Day
- Month
- Geolocation
- Contributor
- “Memory of” (the originator of the memory; not necessarily the contributor)
- Description of the memory
- Social tags
- Attachments (images, videos)
- Terms and Conditions (a compulsory check-box)

The portal supports the submission of JPEG files (3Mb per image), as well as video files in AVI, MP4 and WMV formats (20Mb per video file). Users can submit up to five attachments per memory. With the attachments, the items would then be viewed as a composite memory.

Contributions are published in the portal immediately. The aim was to be as open as possible, relying on the community of users to help flag inappropriate contributions. Behind the scenes, the site administrators would then review published content, to ensure contributions adhere to the Terms & Conditions and remain in true spirit of the project.

It was hoped that through the SMP portal, information gaps can be addressed through the contributions of others. Over time, what the community collectively contributes should help enrich one other’s submissions.

One key premise was that each deposited memory would be linked to at least one other related memory. Contributors would then discover and connect with others who have deposited similar memories. It was hoped that such a feature in the portal would not only reinforce the relevance of adding a memory, but also sustain the rate of online contributions.

4.2 SG Memory Smartphone App

Along with the portal, an iOS app (called SG Memory) was developed and released through the Apple App Store. The app’s focus was to increase the accessibility of the SMP online submission channel. Stories with accompanying images and videos, with geolocation information, could be submitted to the portal on a mobile device. The app also provides a visual way of interacting and viewing memories based on one’s location.

The app captured the same fields as the portal submission form. Submitted memories are directly stored into the same Content Management System, they way they would have been as if the user had used the web portal.

The app was part of the SMP’s positioning that memories could also be “captured as they happen.” In part, it was to address the misconception that the memories only happened in the distant past. By allowing a more immediate means of submission, relatively recent ‘memories’ and experiences could be submitted.

4.3 Twitter.com

The SMP initially adopted the Twitter platform as a means to engage Twitter users. Specifically, the Twitter hashtag #SGmemory was created as a way to uniquely identify tweets that were related to the SMP.

Twitter is known as a “microblogging” platform. Each “tweet” is limited to 140 characters. Accompanying images and geolocation could be attached with the tweet. In March 2012, the SMP experimented with a Twitter campaign, which resulted in the #SGmemory hashtag trending on Twitter over three days, generating more than one million impressions and over 10,000 tweets with #sgmemory hashtag. It became clear that memories could also be expressed as individual tweets. Twitter seemed to be an exceptionally spontaneous channel to generate shared memories.

As a result of the successful campaign, there was a modest steady stream of tweets with the use of the #SGmemory hashtag. This effectively identified tweets as relevant memories for SMP. However, collecting tweets and displaying them required a different approach other than direct online harvesting. The primary consideration was to adhere to Twitter’s Terms of Service and their API rules.

The SMP is currently evaluating options on ways to collect, organise, display and eventually provide access to relevant Tweets. Some of the ways include leveraging on Twitter APIs, or third-party tools, to retrieve and export tweets. Operationalising this would require adherence to the platform’s Terms of use. For instance, Twitter’s Terms of Service states that users retain their rights to any content they post or display on or through their service.⁷ In Twitter’s API terms, developers can use the Twitter APIs to “export or extract non-programmatic, GUI-driven Twitter Content as a PDF or spreadsheet by using “save as” or similar functionality. However, Tweets cannot be exported to a cloud-based service. Developers cannot “sell, rent, lease, sublicense, redistribute, or syndicate access to the Twitter API or Twitter Content to any third party without prior written approval from Twitter.”⁸

4.4 Facebook.com

Facebook is arguably one of the most dominant social media platforms in Singapore. There was an estimated 2.7 million Facebook users in Singapore, or a 76.18% penetration rate of online population.⁹ There are a few popular and active Singapore heritage-related Facebook pages like “On A Little Street In Singapore”, “Heritage Singapore Food”, “Singapore Heritage, Monuments and Places of Interest”, and “Heritage Singapore - Bukit Brown Cemetery”. There are also various school Alumni pages, estimated to number the hundreds.

As with Twitter, Facebook has its own terms of use and API rules. At the writing of this paper, the SMP is developing a Facebook app that will serve as an online submission form. It was very important that the SMP satisfy all legal requirements of the collection process via Facebook.

The app would interface with the SMP content management system, through existing web services as well as within the stated Facebook developer terms of use. The app would function as an online submission form within the Facebook environment. At the same time, the app would acts as a digital pipeline, transferring the same content to the SMP portal for display. Memories that are posted through the app would be deposited and stored directly into the SMP Content Management System.

⁷ “Twitter / Twitter Terms of Service,” Twitter/ Twitter Terms of Service, accessed August 27, 2012, <http://twitter.com/tos>.

⁸ “Developer Rules of the Road | Twitter Developers,” Twitter Developers, accessed July 14, 2012, <https://dev.twitter.com/terms/api-terms>.

⁹ “Singapore Facebook Statistics, Penetration, Demography - Socialbakers,” SocialBakers, accessed August 22, 2012, <http://www.socialbakers.com/facebook-statistics/singapore>.

4.5 Blogs

There are no definitive published statistics on the total number of Singaporean bloggers. A search for Google-hosted blogs (i.e., Blogspot.sg) that originate from Singapore show about 20.4 million results.

From a content collection perspective, blog posts are no different from digital images or any digital memory object, in that they exist in a tangible expressed form. Many blog posts take the form of personal diaries and memoirs. Collecting them would require the same considerations as collecting memory content from individuals or partners, where owners of the content are contacted and their explicit permissions sought.

The primary challenge for the SMP was to identify—on a large scale—blogs and blog posts relevant to the project, and to be able to contact and obtain permissions from the content owners. At the time of this article, the SMP is looking at solutions that include commissioning a online data-mining service combined with a Business Process Outsourcing contractor. The former would identify relevant online content, while the latter would handle the process of contacting blog owners.

Another aspect is the presentation and enabling users to make sense of the collected blogs, with the other memories submitted from various channels. The blogs themselves are the best way to present the context of both the blogger's original expression of intent, as well as make the most sense for readers. Unlike a digital image that has not been uploaded online, blog posts are already publicly available. The approach that SMP has adopted is not to make copies of them by default, but to have the blog owners pledge the posts. Each post, which might be augmented with images and videos, is treated as a composite memory as if submitted through the portal. Instead of displaying the individual blog post in the portal, a link is provided so that users are directed to the blog. The blog posts are counted as memory contributions.

As blog posts are pledged and collected, each blog or blog posts will be reviewed with the purpose of digital preservation. For this, the NLB already has the technology and process for web archiving.

5. The SMP Digital Preservation Plan

Digital preservation involves the processes and operations of ensuring the technical and intellectual survival of digital objects through time. Through the legal deposit law, every Singapore publisher must deposit copies of every publication published in the Republic with the NLB.¹⁰ The NLB has undertaken the task of preserving Singapore's published heritage as part of its responsibility. A review conducted by the Board in 2005 had recommended the building of infrastructure and a centralised database for the preservation and access of Legal Deposit materials and the wider ambit of national heritage materials.¹¹ To this end, the Digital Preservation System was implemented. The NLB is the second national library to implement the Rosetta system by Ex Libris, after the National Library of New Zealand.

Being memories of the nation and directly related to Singapore's heritage, NLB will include selected material from the SMP collection for long-term digital preservation. This becomes necessary over the long-term given that the digital form can be easily changed as time passes, through changes in format, technology

¹⁰ "National Library Board Act," National Library Board Singapore, accessed July 14, 2012, http://www.nlb.gov.sg/Corporate.portal?_nfpb=true&_pageLabel=Corporate_portal_page_aboutnlb&node=corporate%2FAbout+NLB%2FNLB+Act&corpCareerNLBParam=NLB+Act.

¹¹ Chow, Lily and Lim Siew Kim. "The Legal Deposit in Singapore," *CDNLAO Newsletter* 56(July 2006), Special topic: legal deposit system, accessed August 22, 2012, <http://www.ndl.go.jp/en/cdnlaol/newsletter/056/564.html>.

or changes in needs and purpose. There are inherent risks to the digital form that challenge their long-term availability. A digital object also has a shorter life-span compared to a physical item.

The following are potential risks to the life of a digital object:

- Deterioration of the storage medium
- Obsolescence of the storage medium
- Deprecation of format
- Obsolescence of the software
- Obsolescence of the hardware
- Failure to document information about the digital content adequately, and

Long-term management needs The NLB digital preservation system is based on the following international preservation standards:

- Open Archival Information System Reference Model (OAIS-RM)
- PREMIS preservation metadata standard
- METS (Metadata Encoding and Transmission Standard)
- PRONOM (a technical registry for file formats)
- DROID (Digital Record Object Identification)
- JHOVE (JSTOR/Harvard Object Validation Environment)
- PLANETS preservation planning workflow

The Rosetta system enables various processes based on the OAIS reference model to be performed on a content object as it is ingested for digital preservation. This includes integrity checks such as checking the authenticity of the object through checksums for error-detection and the extraction of technical metadata about the object. This is done for validation and to capture technical characteristics of the digital object to be preserved. These information are included as part of preservation metadata. Bypassing details and their in-depth support in digital preservation, it can be said that JHOVE is used to extract technical metadata while DROID validates the file format information for the object based on data from the PRONOM technical registry.

Rosetta largely supports PREMIS preservation metadata requirements and captures these in a METS profile currently under review as a Library of Congress' METS registered profile.¹² This profile encapsulates the expression of data in the AIP for permanent storage. Compliance issues have emerged in the operationalisation of the system by NLB and these have been explored in a separate paper.¹³

Personal memories submitted by individuals and institutions are diverse and composite in nature. The aim is to digitally preserve the memories in the original context in which they were shared, to ensure the continuity of the emotions that might be evoked when someone views a memory in the social, political and cultural environment of the present.

Another consideration is to ensure that the information objects, originally collected as composites, could also be managed as individual digital items for future retrieval and curatorial work. The time span for this is indefinite and can span as much as a hundred years, provided the intellectual rights are available and so long as the technology and the opportunities are available for the preservation and re-use of the digital objects.

¹² "METS Profiles: Metadata Encoding and Transmission Standard (METS) Official Web Site," Library of Congress, accessed August 22, 2012, <http://www.loc.gov/standards/mets/mets-profiles.html>.

¹³ Haliza Jailani and Peter McKinney, "Compliance Conundrums: Implementing PREMIS at two National Libraries," *Archiving* 8(2012): 244-49.

NLB takes the approach of devolving digital content into its most basic form for digital preservation. This means a collection of objects that form an intellectual creation is captured as individual objects. It also means that each object is further broken down into its representations, files and bitstreams. This is in alignment with the PREMIS data model.¹⁴ This approach facilitates long-term preservation action as each object is technically preserved in its most native form. Such an approach reduces risks in preservation action which can be caused by complications in capturing extraneous information to do with the manifestation of the object rather than the intellectual form of the content.

For instance, composite memories comprising of descriptive text, photographs and videos will be captured as separate intellectual entities for preservation. This means they are technically validated and analysed as PDF, JPEG, AVI, MP4 or WMV depending on the type of file submitted. This makes them easier to ingest, validate and technical metadata extracted will be precise. Future preservation action can be performed well when technical information is available in a precise form and the workflows prescribed by PLANETS for preservation planning can be activated.¹⁵

However, when captured as individual objects, there is a critical need to ensure that their provenance and context as found in the memory are equally preserved. Such provenance captures the ownership or custodial history of the digital object and its relationship with the individual or institution whose memory it resides in. These are important in validating its authenticity and the interpretation of its use well into the future.

To ensure traceability for future validation, digital provenance is equally important and the system captures preservation events for each individual object in a composite memory. These include activities such as migration action from a deprecated format to a usable format including the circumstances surrounding such action and the reasons why.

In a similar way, the context surrounding the creation of the object and its relevance to the memory are captured through both descriptive (summaries and abstract) and technical information such as information about the creating application. An image scanned for online submission as a jpeg file is noted as a derivative copy and its provenance and context in the memory are captured.

When a similar image appears in another memory but in its native Tiff, both objects can be linked as representations and the memories connected to form a collective memory. The individual memory is preserved along with the provenance and relevance of the object to each memory. At the same time, they can both be preserved together to form a curatorial coherence within a context that is clear and that will make sense enough to evoke the original emotive reaction as much as possible to future generations of Singaporeans dipping into the Singapore Memory archives.

6. Practical considerations and challenges in SMP digital preservation

From the SMP experience, there tends to be a mistaken sense that what has been digitized is automatically in a “digitally preserved state.” The predominant concern with many contributors tends to be about copyright and intellectual property. The issue of how items are preserved, or whether they would be preserved at all, is often left unquestioned.

¹⁴ PREMIS Editorial Committee, “Data Dictionary Section from PREMIS Data Dictionary for Preservation Metadata, version 2.2,” (July 2012), 7.

¹⁵ “PLANETS: Home,” PLANETS, accessed August 22, 2012, <http://www.planets-project.eu>.

The need for preservation is a common requirement of SMP partners. The main difference lies in the selection approach of the materials to be preserved. For example, universities and research organizations will give importance to the research value of digital content to be preserved. Archives will generally analyse the business activity which will impact the archival value of the records for preservation. While libraries will typically assess the value of its collection that will impact its preservation based on the service to be provided and its institutional mandate.

SMP's approach has been to selectively preserve the 5 million memories that project is expected to achieve. There are pragmatic reasons for costs and resources that influence this decision. What is less clear is the cut-off point of what is an acceptable quantity likely to be preserved. A related question has to deal with the subjectivity of what is worthy of being preserved for the long-term.

For the SMP, developing a selection guideline and policy is perhaps the most challenging of all. Given that the SMP does not wish to be deterministic on what constitutes a "Singapore Narrative" too early at this stage, the selection approach would have to be closely refined.

Appraisal is the first step in assessing a digital object or a digital collection's importance for preservation. It is the process of determining whether the content has permanent archival value and is suitable for preservation. The basis of appraisal decisions may include a number of factors, including the following:

- Provenance and content
- Authenticity and reliability
- Order and completeness
- Condition and costs to preserve
- Intrinsic value

Such appraisal action takes place within NLB's larger institutional collection policies, its legal mandate and organizational objectives.

The prime consideration in identifying and selecting digital content for preservation is that the content must be about Singapore or be of interest to Singaporeans. This will result in a huge data set. There will always be limited resources to support digital preservation in the face of the amount of digital content available. To this end, the need to appraise, select and prioritise among digital content for preservation becomes imperative.

There exists a challenge in having to reconcile the non-standardised and composite materials submitted by individuals and institutions. Another level of complexity is added when such collection and coordination efforts are happening concurrently and at multiple fronts.

Tradeoffs and challenges are to be expected as the SMP progresses. Memories may not always be complete. In trying to seek a balance between web site usability (this includes benchmarking the site with popular social media platforms) and the needs of item documentation, there are inevitable compromises.

For instance, there are only three compulsory fields or actions when users submit a memory. The description and intellectual rights information are not mandatory. When submitting their memories, most users do not include details in the description box such as information about the photographer, videographer, copyright holder, license details or other descriptive contextual information which would stand the memory in good stead for future preservation.

At present, memories that have been submitted cannot be edited or removed by users. This is to ensure the integrity of the memory submitted is maintained. In practice however, several users have requested for an editing feature merely to make minor amendments to grammar, or to replace a wrongly submitted image. Technically there are solutions possible such as keeping versions of the memories

submitted, but this leads the team to question whether the system will be an over-engineered one which can lead to a higher-level of maintenance and increased costs.

The authenticity of the submitter is not necessarily verifiable. To keep the online platform as open as possible, our design adopted the social login features for popular sites like Facebook, Google, Yahoo and MSN. This was a welcome feature by users, as it meant they did not need to manage yet another user account. But being ‘open’ also meant that the same user can submit similar memories under different accounts. The overriding principle for SMP is that the core requirement of a memory would be the content itself. While the originator of the memory provides useful contextual information, the SMP would need other ways to verify the authenticity of the content.

That said, there are these guidelines for the selection of SMP content for digital preservation:

- Content with authoritative information of national significance (e.g., information from government agencies), are of historical or research value or which contain authoritative information of historical, scientific, social, political and cultural significance from other national agencies such as educational institutions and societies.
- Content which target a primarily Singaporean audience published locally or overseas.
- Content which feature Singapore, or are of interest to Singaporeans, i.e., such content may originate from other countries
- Content of events with national impact that take place in Singapore, which may be topical in nature.
- Content about prominent personalities in Singapore or content published by individuals who are authoritative and knowledgeable in their fields, whose views and works are popular in public opinion or provide a good social commentary.

It could be possible that the SMP to use the approach of crowdsourcing, where people—particularly online users—could help to directly or indirectly identify posts worthy of long-term preservation. For example, every memory item displayed could have accompanying checkboxes for users to rate: “most interesting”, “more people should read about this”, “I know others who might connect with this”, “worth preserving”, etc.

In the case of Tweets, although SMP cannot directly copy and display Tweets with the #SGmemory hashtag due to Twitter’s terms of use, selected tweets deemed significant to the collective national memory might potentially be downloaded in CSV format and preserved as dark archives. This preservation decision could be aligned with the responsibility that NLB has undertaken to preserve the memory of the nation.

Technically the CSV format is more amenable to digital preservation as it is a native form that is more open than the proprietary Twitter software. Although it will not retain the original look and feel of the Twitter posts, the data is already devolved to the level of the raw intellectual content which can be preserved easily.

Once the intellectual content has passed into public domain, it will even be a better option to keep the CSV form of the raw content as it is an open format that can be used easily in computer systems, with the possibility of feeding it to a future viewer that may be using new technology unthought-of in our current time.

On the whole, the SMP will continue to take pragmatic steps in balancing the immediate need for collecting memories and the longer term needs of digital preservation. Digital preservation policies and processes would be incrementally implemented as clarity increases with the amount of memories collected.

7. Conclusion

The goal of the Singapore Memory Project is to engage individuals, communities, groups or institutions who have formed memories and content about Singapore and would like to contribute them. This will build a culture of remembering which will nurture bonding and rootedness. It is envisioned that the project will continue so long as there are contributions from the community.

Apart from a quantifiable target, the SMP has intangible aims of getting as many people as possible taking part in the history of their family, streets, country and world. There is a larger aim of getting people from different generations talking more, sharing more and coming together more often. Ultimately, like all memory projects perhaps, the SMP aspires to be an open global archive for everyone to enjoy, learn from, use and reflect.

It is also important that people can make sense out of the memories that are (and would be) collected. The on-going challenge is also on how to collectively presenting the content collected from the portal and various social media channels.

The term “digitization” often carries with it an implied understanding, albeit inaccurately, to mean the same as “preservation for future generations.” Memory projects like the SMP would be good opportunities to educate the wider public about the concept and need for digital preservation.

Preserving our documentary heritage in digital form helps the nation to nurture a sense of belonging and national identity. By preserving these knowledge assets, libraries help to sustain their value in bridging information and promoting learning. This would enable libraries to engage new audiences, support innovation and research as well as foster sharing and learning as the nation grows.

National Libraries are arguably well-placed to provide such a shared service, in addition to being the lead agency in developing policies, guidelines, standards as well as public and stakeholder education. Future generations will be able to depend on the library to provide well-preserved digital documentation of information and knowledge as they are created, shared and built upon and events as they happened, and be able to use these valuable artefacts which would be lost otherwise.

Digital Preservation Policy of The Chamber of Deputies¹

Methodology for its development

Ernesto C. Bodê

Abstract

It's about the practical methodology used for writing the Digital Preservation Policy of the Brazilian Chamber of Deputies. The methodology uses the principles of work structuring, using macro steps to achieve the final text of the policy. In this work, we aim to derive and present the methodology in a way that might be applied by any other institution. To illustrate the use and application of the methodology, we present how the real work was done in the project. The main final products of the proposed methodology are also presented.

Author

Currently a doctoral student in information science (University of Brasilia - Brazil), with a Master's degree in Information Science (University of Brasilia - Brazil). Graduated in Archival and Library Science (University of Brasilia - Brazil). Analyst in the Centre of Documentation of the Chamber of Deputies in Brazil. Was the Project Manager for the Digital Preservation Policy of that institution.

1. Introduction

The objective of this paper is to present a practical methodology for the creation of an *Institutional Digital Preservation Policy*. This methodology was derived from our own experience in a project for the creation of a digital preservation policy in *The Chamber of Deputies* (National Congress of Brazil). First we present an overview of the methodology and the necessary stages proposed for implementing it. For each stage we present the goal and the general idea inside it. In addition, each stage contains a description of the way that effectively performed the proposed procedures. For each stage, we also present the real products produced in the *Chamber of Deputies*, i.e., reports, studies, up to the final product which is the final text of the policy. The policy that originated the execution of our work is in final form adopted by *The Chamber of Deputies*. Before discussing the methodology itself, we shortly present some concepts used throughout the paper.

2. Methodologies and Institutional Policies

Professionals will always devise ways to perform their regular work tasks. However, the pace of work and speed requirements often dictate how tasks should be performed, regardless of whether or not this is the best possible way to perform such tasks. Careful planning and the use of an adequate methodology, like the one described here, will help to mitigate this problem. Another advantage of using previously designed methods is that this will make it possible to promote further improvements in the procedures.

¹ The Brazilian Parliament is called *National Congress*. Brazil has a bicameral legislative assembly, composed by *The Chamber of Deputies* and *The Federal Senate*.

In this work, a methodology is understood as “A set of guidelines or principles that can be tailored and applied to a specific situation.”²

We are dealing here with institutional policies, i.e., a set of high level guidelines, applicable to an organization. In this article, the concept of policy is understood as follows: “*predetermined course of action established as a guide toward accepted business strategies and objectives.*”³

3. The Methodology Used

The use of previously *tested* and *approved* methodologies for the implementation of any project is always an approach that leads to better qualitative and quantitative results, e.g., reduction in the total number of hours spent in the project.

Like many institutions worldwide, *The Chamber of Deputies*,⁴ has felt the impact of the use of digital documents and technologies on their work processes. Therefore, there is a clear need for a digital preservation policy that provides greater security for the collections of digital documents. However, regarding projects aimed to assist in the implementation of policies, and more specifically, digital preservation policies, no tested practical methodologies are publicly available for use and for the development of a project like the one developed in our institution. In this scenario we had to propose and adopt a new methodology of working. Using this methodology allowed it to be tested and improved. In fact, our current knowledge on this now tested methodology has only been possible after the completion of all the stages that culminated with the official version of our *Digital Preservation Policy*.

We highlight two benefits of the use of the proposed methodological approach, besides the advantage of having already been tested with *actual available products*, that is, with a *Working Model* to guide the activities' path. First benefit: reduction in **subjectivity**. Policy texts are always elaborated by people, that is, they are products of human intellect. Hierarchical and personal differences within a team can cause some opinions to prevail over others. Also, besides the misunderstanding that may arise from this process, the views that prevail are not always the most appropriate, even if they are expressed by experts. Anyway, some degree of subjectivity will always exist, but we believe that the use of the methodology proposed here will help to mitigate this problem.

Second benefit: parallel to the problem of subjectivity is the need for theoretical **argument** for the guidelines in the policy text. Through the use of the methodology, each guideline is well-founded, and, consequently, the whole internal discussion and persuasion process tends to be smoother. Again, as it is something produced by people and as part of institutional negotiation processes, some guidelines may exist that are obtained through an institutional negotiation process rather than through the use of technical reasons. However, this problem is also expected to be mitigated. The reasons we support these two benefits will be clarified after the presentation of the methodology itself.

The methodology proposed here is not complex. It is rather an eight-stage methodological work, each one delivering an **actual product** (presented and exemplified below). The last product is the official version of the policy adopted by the institution.

² Jason Charvat, *Project management methodologies: selecting, implementing, and supporting methodologies and processes for projects* (New Jersey: John Wiley & Sons, 2003), 3.

³ Stephen B. Page, *Establishing a system of policies and procedures* (Ohio: BookMasters, 1998), 2.

⁴ <http://www2.camara.gov.br/english>.

The methodology is grounded on the **Technical Reasons** for proposing guidelines for drafting a **Digital Preservation Policy**. This reasoning shall be based on an *Initial Source* of information on *Digital Preservation* and the establishment of a *Areas* for the policy text. The *Initial Source* is the basis for the construction of the *Areas*. As soon as the *Areas* are available, the other steps of the process take place with the guidelines being drafted before the final official version is obtained and ultimately approved. Also, a step for the procedures of internal discussion and persuasion process regarding the policy text is proposed.

As part of the methodology, the adoption of consolidated technical standards is suggested, not only regarding *Digital Preservation*,⁵ but also *IT Governance*,⁶ *Repositories*⁷ and *IT Security*.⁸ In addition, previously defined technical terms must be included, and a *Glossary* of the terminology used in the policy text must be provided, which shall contain all of these technical terms. These technical words will be gradually added to the *Glossary* throughout the eight-stage process and the latter will finally be added to the final product as an *appendix*. The use of a *Glossary* of terminology based on widely known technical terms avoids lengthy discussions on the concepts used in the policy.

The search for institutional support to develop the project in the case of *The Chamber of Deputies* and the selection of the team responsible for project management up to the step of discussion with the other members is not listed here as a stage in the policy development process, though it is a key “pre-processing step.”

3.1 The Stages of the Methodology

The methodology developed and used in the elaboration of the *Digital Preservation Policy* in *The Chamber of Deputies* involves an eight-stage process:

- Original Source
- Areas
- Pre-guidelines
- Minute Drafting
- Internal Persuasion
- Official Formal Writing
- Official Approval
- Revision

3.1.1 Original Source

Overview

Each new policy to be developed includes a specific domain of technical knowledge related to the main theme addressed. In this step of the process of elaboration of the policy, the project team should collect all the technical information available on **Digital Preservation**, **Institutional Policies** and **Digital Preservation Policies** to support the project. It is desirable that at least one member of the team has expertise in *Digital Preservation*. Alternatively, support can be sought from external consultancy.

⁵ See ISO/TC 46/SC 11 Digital Records Preservation.

⁶ See ABNT NBR ISO/IEC 38500:2009.

⁷ See ISO 14721:2003.

⁸ See COBIT 4.1.

The source of this particular information can be scientific papers, institutions that developed similar projects and experts on topics of interest. *Digital Preservation Policies* already done or similar works from institutions that have already developed such projects are an excellent source of reference for the proposed activities.

How we did it

With respect to *The Chamber of Deputies*, the team has always relied on the assistance of a *Digital Preservation* expert. Thus, the necessary expertise has been naturally incorporated into the project. On the other hand, as we explained above, the specific issue of *Digital Preservation Policies* has been little explored and not much information has been made available to the public. The alternative adopted in this stage was the development of a *benchmark* for policies or similar strategies elaborated or adopted by other institutions.

After a preliminary study with technical definitions on the *basics* of digital preservation, we carried out a survey and identified thirteen policies or similar texts available in the Internet to support our *benchmarking* activities. These texts are listed in Table 1.

Table 1. References for the original source.

INSTITUTION	COUNTRY	TITLE OF THE WORK
Direção Geral de Arquivos, Ministério da Cultura	Portugal	Repositório de objetos digitais autênticos: política de preservação digital
National Library of Australia	Australia	Digital Preservation Policy
UK Data Archive	United Kingdom	UK Data Archive Preservation Policy
National Library of Wales	United Kingdom	Digital Preservation Policy and Strategy
Online Computer Library Center	USA	OCLC Digital Archive Preservation Policy and Supporting Documentation
University of Minnesota	USA	University Digital Conservancy Preservation Policy
Yale University Library	USA	Digital Preservation Policy
Arts and Humanities Data Service	United Kingdom	Collections Preservation Policy
Columbia University Libraries	USA	Policy for Preservation of Digital Resources
British Library	United Kingdom	BL Digital Preservation Strategy
State Library of Victoria	Australia	Digital Preservation Policy
Inter University consortium for political and social research	USA	ICPSR Digital Preservation Policy Framework
Cornell University Library	USA	Cornell University Library Digital Preservation Policy Framework

Note: All documents available online Mar/2010.

An analysis of each source identified in the *table 1* has been made to highlight the most important points and the standards adopted.

Products delivered

Report 1 that includes a preliminary study on concepts of digital preservation and *benchmark work* with thirteen institutions previously identified and assessed.

3.1.2 Areas

Overview

The specific knowledge on *Digital Preservation* and research on previous policies or similar strategies is aimed to enable, in the second stage of the process, the development of *Areas* for the policy text.

One key principle for the development of a methodology is decision making based on pre-established and consistent criteria. In *stage 1*, the purpose of the research on knowledge about digital preservation approaches was meeting the referred principle. The development of *Areas* for the policy text will allow the pursuit of the project based on criteria aimed to mitigate the inherent subjectivity of a work team, even in the final steps where other actors will participate in the policy refinement process.

The name and purpose of each *Area* are defined based on information provided in *stage 1*. Another advantage of the use of *Areas* is that it makes it possible to streamline working procedures so that team members are assigned different areas.

How we did it

The study of *Report 1* project led to the establishment of eight⁹ *Areas* to be contemplated in a *Digital Preservation Policy*. There is no exact match between these eight areas and the parts of the policy text. What matters here is to establish all the key elements to be contemplated.

1st Area: Aims of the policy and its establishment

It is about the purposes of the institution that elaborates this policy and clarification of the relationship between these aims or objectives with the policy currently under construction.

2nd Area: Objectives of the policy

It is about all the aims or objectives of the policy currently under construction. The following items specify guidelines for action, but the more general aims and objectives should be recorded in this part of the policy text.

3rd Area: Relationship with other policies

It is about the relationship of the text of this policy with other policies or strategies of the institution, or else with available legislation related to the objectives and actions set out in this policy. Specifically mention the standards to which the policy complies and upon which it is based.

4th Area: Scope of the policy

It is about the scope of the digital documents comprised by the policy. Non-scope can also be stated. Another possibility is to establish the priority for action in relation to the groups of digital documents within the defined scope, thus, prioritizing some digital documents over others. This is a solution to the problem of accepting a wide scope and, at the same time, prioritizing some groups of documents more relevant to the institution. When priorities for action are defined, the levels of support for some file formats are also defined. However, because of their importance, file formats should be addressed separately.

5th Area: What and how to preserve

It is about what the policy is committed to preserving in digital documents (specifying the documents) and for how long, e.g., preserving only the bit sequences, access to content and/or its relationship to the social context in which it was created. The commitments made directly affect the complexity of the

⁹ Different projects in other institutions could derive more or less than eight areas.

actions to be implemented to comply with the referred commitments. Therefore, attention should be paid to the possible scenarios before accepting a certain level of commitment.

6th Area: Cooperation with other institutions

It is about declaring that the institution cooperates or intends to cooperate with other institutions. This element needs special attention, for it is necessary to ensure that the institution really intends to establish ties with other institutions. It could also be introduced in future versions of the policy.

7th Area: Roles and responsibilities

It is about the roles and responsibilities undertaken by the different bodies and offices of the members of the organization regarding *Digital Preservation*. This element will be the product of the study of all the actions set by the policy text in its relation to the different bodies and offices of the members within the organization.

With respect to the responsibilities undertaken, pay attention to the *Open Archival Information System* (OAIS).

8th Area: Governance

For this group, perhaps more important than defining what to do and how to do it, is establishing guidelines on the periodical updating of these recommendations. For example, the establishment of an *Internal Management Committee*¹⁰ for periodically reviewing the policy text and suggesting changes to the subsequent versions of the text. The content of this element involves recommendations of specific accepted and recommended Technologies, as well as the necessary procedures, the most important items include:

1. Document and track changes made to formats, media or any other technology used during the custody of digital documents;
2. Who will evaluate the file formats recommended for use and specific applications such as still image, video or sound;
3. Who will evaluate the specific media to be used in the institution and how it will be done;
4. Define the procedures and strategies to be adopted, e.g., the migration of file formats and media;
5. Define the authorized file formats (recommended, mandatory) and where they should be used. Define also the versions of the authorized file formats.

Products delivered

Report 2 with the *Areas* of the **Digital Preservation Policy**, specifying which areas and objectives will be contemplated.

3.1.3 Pre-guidelines

Overview

In this stage, **pre-guidelines** will be introduced in each of the *Areas* previously established. The name of the *Area* and its objective are the main source for the selection of these **pre-guidelines**. We use here the

¹⁰ Another policy in *The Chamber of Deputies* created this committee. Chamber of Deputies, “Digital Preservation Policy,” [in Portuguese], 2012, <http://www2.camara.gov.br/login/int/atomes/2012/atodamesa-48-16-julho-2012-773828-norma-cd.html>.

term “pre-guidelines”, because this stage is not aimed to create the definitive version of the guidelines of the policy. Our concern here is to insert in the policy all the important guidelines related to each *Area*. The adequate **elaboration** and **drafting** of these guidelines are further steps in the process.

One advantage of this methodological approach is that, at this point, it is possible to streamline working procedures by assigning the tasks of insertion of **pre-guidelines** according to the area of the team member. Thus, there may be focus of some members on specific areas, increasing the possibility of detection of all the **pre-guidelines** necessary for the institution considered.

All possible **pre-guidelines** should be introduced in this step. The refinement process, that is, the possible removal of unnecessary or redundant **pre-guidelines** (relative to other areas of the framework) will take place in subsequent steps.

What is the source for the insertion of these **pre-guidelines**? As already mentioned, the first source is the objective of the area where it is being inserted. Other sources should be sought by the team members responsible for populating the respective area, such as papers¹¹ on digital preservation, policies or similar strategies from other institutions.¹²

How we did it

Tasks were assigned to team members for different *Areas*. Proposals of pre-guidelines related to the objectives of a given area were made.

Below is an illustrative example of proposal made for 3th Area (*Relationship with other policies*):

3.1

This policy attempts to meet the Brazilian standard 15.472 of April 09, 2007 in its reference model for an open archival information system (OAIS).

Rationale for this pre-guideline development

One major assumption in the policy development process was the use of standards, especially the OAIS standard, which establishes the information model and the minimum responsibilities that should be taken on to achieve successful digital preservation.

Products delivered

Report 3, which includes the first proposal of **pre-guidelines** by *Area* of the policy.

3.1.4 Drafting

This process step is aimed to transforming the content of the pre-guidelines, which has been previously proposed and defined, into a clearer and direct text.¹³ In this step, besides the use of the vernacular, another concern is the arrangement of the guidelines in hierarchical order, that is, more comprehensive guidelines should precede less comprehensive guidelines. It is necessary to eliminate redundancies

¹¹ Neil Beagrie, Najla Semple, Peter Williams, and Richard Wright, “Digital Preservation Policies Study,” Part 1: Final Report October 2008, Prepared by Charles Beagrie, Ltd., Funded by JISC, http://www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport.pdf.

¹² Luciana Duranti, Jim Suderman, and Malcolm Todd, “A framework of principles for the development of policies, strategies and standards for the long-term preservation of digital records,” InterPARES 2 Project, 2008, [http://www.interpares.org/public_documents/ip2\(pub\)policy_framework_document.pdf](http://www.interpares.org/public_documents/ip2(pub)policy_framework_document.pdf).

¹³ Michael R. Overly, *e-policy How to develop computer, e-mail, and internet guidelines to protect your company and its assets* (USA: Amacon, 1999), 4.

between the different pre-guideline proposals submitted. This step also involves detailed examination of the wording of each pre-guideline to obviate doubts regarding its use.

Terminology is regarded as important throughout the entire process, but will deserve further consideration from this step on (fourth stage), which is precisely focused on drafting *guideline text*. The use of consolidated terms from standards can save a lot of time and effort regarding the best way to draft the text of guidelines.

How we did it

We read all versions submitted for each contemplated area. We began a drafting process that included several text versions. Guidelines were grouped by chapters¹⁴ to be addressed.

After several revisions, the first version, but not the final one, of one article of the policy reads as follows:

- Art. 4 The objectives of the Digital Preservation Policy of The Chamber of Deputies are:*
- I. ensure the right conditions to provide full-text access to all digital documents for the period of time specified by the institution;*
 - II. create (adopt) standards, instruments and mechanisms to ensure the permanent authenticity of digital documents;*
 - III. define the software and hardware requirements for the establishment of the institution's own digital repository to allow for digital preservation;*
 - IV. contribute to reducing information security risks;*
 - V. promote the exchange of information and experiences on digital preservation with national and international institutions, aiming at reducing the effort and cost involved in finding solutions;*
 - VI. disclose [train?] and share knowledge and best practices of digital preservation;*

Products delivered

The product delivered in this step was the first version (draft) of the **Digital Preservation Policy** of *The Chamber of Deputies*.

3.1.5 Internal persuasion

This step requires the existence of a basic text (a first version) of the **Digital Preservation Policy**. The text obtained here is not necessarily the final version of the policy, but is supposed to contain the essential proposal. After the preparation of this text, it should be checked and, if appropriate, adjusted to the needs and limits of the different bodies within the institution.¹⁵

This step is essential because different actors will contribute to the success of the policy within an institution. Thanks to the commitment and partnership of those actors, the chances of success of the final product will increase considerably. On the other hand, the actual scenario of each institution should be taken into consideration. When it comes to people, the subjectivity of each of them must not be neglected, so that situations where it may be necessary to impose the policy or overcome resistance to its implementation must be considered for each specific case. Some issues may be just a matter of lack of communication, that is, briefings and meetings to present the policy can not only clarify doubts, but also

¹⁴ These “Chapters” do not have an exact correspondence with the *Areas* proposed.

¹⁵ Notably, those bodies that will be directly affected by the text of the policy.

gain supporters of the policy. In an ideal scenario, the policy is elaborated by a team composed of people from key bodies affected by the policy. With the support of such a team, the persuasion process will be easier. However, the mere existence of such team does not necessarily mean support to the text submitted to the internal bodies.

Another reason for conducting this step is the possibility of improvement or even correction of the policy text, not only in respect to its form, but also regarding aspects not discussed during the establishment of the first guidelines.

In the internal persuasion process, the existence of the *Areas* for the policy text and its respective guidelines, including that reasons for proposing such guidelines, that is, the first steps of the process can and should be used as arguments to make clear that the proposed text is well-grounded and consistent, and is not merely a subjective proposal of the work team.

How we did it

This step was performed at *The Chamber of Deputies* in various ways. First the policy text was made available for internal public consultation, so that various bodies might be informed of it and proposes modifications to the text.

There have also been several meetings with different internal bodies of the institution to present and discuss the policy text. In our case, there has been a partnership between the body responsible for documentation and the one in charge of information technology at the institution. We set a *deadline* for the internal bodies to present their considerations and suggestions relative to the policy text, as well as for clarification meetings.

Finally, to facilitate the internal persuasion process, several documents detailing the policy architecture, that is, how it is structured, were produced. We also produced a document with comments on each article of the policy text. As a result of this process, improvements have been made to the text and new versions were produced.

Products delivered

The products delivered in this step were a new version of the *Policy Text* (with many improvements), a *Policy Text* with comments on each article and a report detailing the *Policy Architecture* (areas covered by the policy and scope of digital documents).

3.1.6 Official Formal Writing (Minute)

Depending on the institution, this step may or may not be necessary. For some institutions the policy text must follow some standard patterns, so they provide models and standards for this type of document.¹⁶

Depending on the institution, it may be recommended that the official formal writing step precedes the internal persuasion step. Still, we believe that it should take place after stage 5 because it is essentially a work of text revision when required.

At any rate, with respect to the work methodology, this step should be done separately from the drafting step, because in the latter focus should be given to content.

¹⁶ Stephen Page, *7 steps to better written policies and procedures* (Ohio: Process Improvement Publishing, 2001), 9.

How we did it

As some team members were also knowledgeable in formal writing, several aspects related to formal writing style were used. In the end there was little need for adjustments in this regard.

Products delivered

The product delivered in this step was an improved version of the *Policy Text*, in the formal writing style. This version was called *Minute*,¹⁷ to which we attached a *Glossary* with terms used in the Policy.\

3.1.7 Official Approval

As in the case of *stage 6*, this step may or not be necessary, depending on the specific institution where the work is being developed.

The final approval mentioned here indicates that the institution's top level of hierarchy is to approve the text. Such official support will be important in the subsequent steps of policy development, that is, the implementation of the procedures and actions to carry out digital preservation. In several institutions, this final approval implies the existence of an internal official standard, which will be specific to each institution considered.

How we did it

This step was performed by the directors of the centers responsible for *Documentation* and *Information Technology*. Negotiations were held at senior management level.

Products delivered

At the end of this step, the product delivered was the text of the *Digital Preservation Policy* as an *internal standard* of *The Chamber of Deputies*.¹⁸

3.1.8 Revision

It may seem odd to include a stage in the digital preservation policy after the final approval step. In fact, this stage is not part of the methodology, but was introduced here because of its importance. Institutional scenarios change, and information technology and computing issues, in particular, are in a continuous state of evolution. Both software and hardware may undergo major changes in little time.

This is not about considering whether there will be technological changes, but to monitor those changes as they occur. The success and effectiveness of the policy text depends on this monitoring and on the actions necessary to adapt the policy text to the new existing scenarios.

How we did it

There was no need so far to revise the standard, but this responsibility is foreseen in the approved official version.

Products delivered

No products delivered.

¹⁷ Here it means a version before the formal and legal one.

¹⁸ Chamber of Deputies, "Digital Preservation Policy."

4. Conclusion

In the limits of this paper we did not present an in depth detailed steps of real work done. But from the use of the practical methodology presented in this paper, we highlight some points. First of all, the use of this methodology allowed the production of a text coherent and well-reasoned. Finally, the conduct of work among team members has been rationalized, i.e., it was possible the division of tasks and consequent focus on specific points.

The work presented here is certainly not finished. We presented here a start point to be improved by others. At least, that's what we hope. It is very important to say that this work could not be possible without so many important research centers and people dedicated to study and improve this new and fascinate area, *Digital Preservation*. The few citations presented here are the ones essential to explain the methodology; many others were omitted due to the limits and format of this paper.

Finally, the main product produced by this methodology became more coherent and well-reasoned. This allows future actions of *Digital Preservation* with higher chances of success in *The Chamber of Deputies*.

Web 2.0 Products as
Documentary Digital Heritage:
Can We Access and Preserve Them?

Unprotected Memory

User-Generated Content and the Unintentional Archive

Jamie Schleser

American University, USA

Abstract

While much discussion of online archives focuses on traditional cultural heritage institutions, a phenomenal amount of original, creative cultural production in the form of user-generated content is being stored on the servers of for-profit corporations like YouTube. These unintentional archives aggregate content in the moment with no concern for preservation. The unique ecology of the corporate online space further complicates the issue of digital heritage as these sites depend on market economic factors like maximizing profitability and minimizing copyright infringement liability rather than ensuring the accessibility of these cultural products for future generations. This paper will analyse relevant policy issues as they relate to the United States specifically, to the transnational domain of both corporations and digital heritage, and to the site-specific codes of conduct that govern these unintentional archives. It will assess emergent digital preservation initiatives in order to present recommendations for future best practices.

Author

Jamie Schleser is a Ph.D. Fellow in the School of Communication at American University in Washington, D.C. Her current research focuses on assessing the impact of the Internet, social media, and mobile technologies on cultural memory, historical narrative, and individual identity. Schleser is particularly interested in the expansion of access and the participatory culture facilitated by the digitization of physical archives and the aggregation of born-digital cultural production, whether formally or informally, into omnipresent web archives.

1. Introduction

When tragedy occurs, communities respond. While much of the focus is often on rallying around those most affected by what has happened, others in the community who have not experienced a personal loss or injury can still be left searching for answers and solace. The process of documenting and digesting these events collectively through communication lies at the heart of cultural memory. In many cultures, the untimely loss of life has long been marked by the gathering of groups and the construction of tangible memorials. Memorialization might be fleeting, in the form of flowers, personal tokens, and religious iconography left at the site of a roadside accident, or relatively permanent, in the form of architectural monuments constructed to mark a significant place. As mediated communication has become the norm, physicality is no longer a prerequisite. With each new connective technology, beginning with the invention of the printing press in the 1400s and followed in the past two centuries by advent of the telephone, broadcasting tools, and the Internet, the scope and size of cultural communication has grown exponentially. While the power of interpersonal communication has not waned, the ritual of communicating through mass media formats described by James Carey has come to occupy a central role in the exchange of ideas and the creation of culture today.

In developed countries, it has become nearly impossible to conduct the business of mundane life—whether connecting with friends and family or participating in the Habermasian public sphere of political

deliberation and opinion formation—without encountering mediation. When a gunman opened fire on the campus of Virginia Tech University in Blacksburg on April 16, 2007, the world was watching. The rampage, in which the gunman killed 32 people and injured dozens more before turning his weapon on himself, was the deadliest attack on a college campus in U.S. history and news of it quickly spread through mass media, receiving coverage nationally and internationally. In the aftermath, students, staff, faculty, and alumni gathered for memorial services on the campus, holding candlelight vigils and creating informal memorials at the sites of the dormitory and classroom building where the attacks took place, but the mnemonic discourse surrounding the tragedy that had occurred there expanded far beyond the local community. Due to mediation, broader awareness breeds a broader response. Individuals without physical or personal ties to local tragedy are now not only exposed to those events, but are also able to construct and share their own personal tributes to its victims. This is particularly true when the circumstances resonate across borders, as they did in the case of Virginia Tech, which struck a raw nerve for many Americans as colleges are often assumed to be sanctuaries of learning for young adults bursting with yet unrealized promise.

Dozens of Virginia Tech memorial videos have been posted to YouTube, a content-sharing platform that allows registered users to upload unlimited material for free. One user, identified only as “auntmannys,” posted such a video within three days of the attacks accompanied by the following explanation: “though we did not know the victims, our thoughts and prayers are with their families” . Little background information about the user is provided, though the memorial is posted alongside other videos created by the same individual, including compilations of family footage and impromptu recordings of living room dance-offs and video game duels. Just over six minutes in length, the video consists of opening text providing some facts about the shooting and an appeal for unity in remembrance of those lost, followed by a montage of photos of the victims, including information about each person’s age, hometown, and role on campus. All of this is set to music reflecting the somber tone and purpose of the video. Other Virginia Tech memorial videos posted to YouTube follow a similar formula, though some focus more on particular victims and are clearly the expressions of individuals who knew them personally. While the video posted by “auntmannys” has received over 1,300 views, a fairly large audience for something created by a pseudonymous nonprofessional who has admittedly little personal relationship to the event other than generalized empathy in the face of human suffering, comparable projects have been viewed by upwards of 40,000 or even 100,000 visitors. This digital form of memorialization, embodied in the content created and broadcast by a distant individual user using a popular corporate content-sharing platform, represents a unique and unprecedented mediated form of cultural memory. If properly preserved, these videos will document a new dimension of cultural response to tragedy that arose after the turn of the millennium, supplementing the often impersonal record of news coverage and official reports, just as material artistic expression and archived personal correspondence have marked this process in the past.

While the remnants of tangible memorials are assumed public and are often cared for by traditional cultural heritage institutions, including museums, libraries, archives, and universities, this new aspect of mnemonic discourse is taking place on private platforms like YouTube, Vimeo, Photobucket, and Flickr. In this corporate domain, cultural communication is bound by site policies that determine what materials are transmitted freely, what speech is censored, what ownership rights individuals maintain when posting user-generated content, and what happens to content once posted. The nature of governance is determined exclusively by the organization providing the service and conveyed to users via jargon-dense Terms of Service (TOS) agreements under the default assumption that participation equals consent. At the same

time, the growing significance of user-generated content aggregated on these platforms—like the memorial video created by “auntmannys”—as a record of cultural memory has yet to be embraced fully by archivists and historians and is even further from the minds of the general public and the owners of these commercial enterprises. While historical practice and cultural scholarship depend on the ability to reference collected artefacts, born-digital modes of communication have magnified both the amount of material potentially in need of archiving and the challenge of fixing those materials for preservation due to their ephemeral and evolving nature. Failing to embrace the rich stores of born-digital cultural production located within these unintentional archival spaces, where material is collected without curation or the intent of preservation and located outside the conventional domain of traditional cultural heritage institutions, could prevent future generations from being able to accurately reflect on everyday life as we know it now. Without a clear understanding of the impact of the privatization of the public sphere as a result of the dominance of these corporatized platforms, there is potential to create an unprecedented dark age of digital cultural memory.

With the goal of drawing attention to the existence of unintentional archives of important cultural artefacts on content-sharing platforms and developing strategies for ensuring the preservation of these resources, I will begin by sketching a broad understanding of digital archival practice and the particular impact of born-digital communication formats. While the most obvious challenge for digital archiving initiatives is the quantity and instability of the material to be preserved, these projects are also affected by changes in how we communicate. As cultural production and archival practice move online, I will discuss how the ecology of the Internet as a space where traditional nation-state borders are less restrictive to the flow of information has resulted in the addition of a globalized mnemonic discourse that must now be negotiated alongside local, regional, and national narratives of the past. Next, I will explore how the movement of personal expression to digital formats controlled by transnational corporations has fundamentally altered the process of archiving cultural memory by subjugating Constitutional directives encouraging the free exchange of ideas for the higher purpose of nurturing public discourse to localized codes of conduct driven by commercial imperatives. As it one of the most dominant U.S.-based transnational content-sharing platforms where an unintentional archive of user-generated cultural production has formed, I will use YouTube as a case study to identify relevant U.S. government policies that influence the business model of these services. I will also examine the body of self-issued laws governing this private platform, the TOS agreement, to highlight consequences of this shift in power from governments to corporations. Finally, I will analyse three experimental strategies currently being used to stimulate the archival preservation of born-digital cultural production in order to make recommendations for future best practices.

2. Digital Archives and Global Memory: A New Paradigm

While archival practice is increasingly turning toward digital formats, there are two distinct approaches being used today. The first, undertaken primarily by traditional cultural heritage institutions, has to do with creating digital copies of existing physical artefacts with the goal of either retaining damaged items that have become impossible to maintain in their original format or expanding access to particular collections for those unable to invest the time and resources necessary to travel to the museum, archive, library, or university that houses them. The second set of initiatives focuses on capturing the wealth of born-digital cultural production now possible thanks to advances in computing and Internet Communication Technologies (ICT) during the last forty years. As everything from broadcast television

to interpersonal communication has moved from analogue to digital, methods for capturing and preserving those materials must constantly play catch-up. The magnitude of digital communications can be explained by the compression of the time and cost necessary to distribute them. Emails convey correspondence instantly and without additional cost for individuals with high-speed Internet access where letters once took weeks to travel to their destination and required postage to get them there. Legacy news outlets can reach hundreds of thousands of readers, often through a single article, where entire newspapers previously had to be printed and delivered to individuals who subscribed or sought them out on street corners. Similarly, the accelerated evolution of ICT has resulted in a constant flow of buzzworthy projects, each one holding the title of “next big thing,” as Napster, MySpace, or the original Friendster (now attempting to reinvent itself as a gaming social network) all did, before being surpassed by something newer and more exciting. The logistical problems stemming from data quantity and the persistent cycle of technological obsolescence are the biggest obstacle to developing a feasible preservation strategy for born-digital cultural production.

The problem of scale is fundamental, as Mike Featherstone has argued that digitization requires that the very nature of the archive be reimagined as a monolith database rather than the ordered physical space of the past. As born-digital cultural production and Internet-based archival practice take hold, there is real danger of being swept up in utopian visions of cataloguing everything and providing universal access. This impulse must be balanced by a rational assessment of the instability of the online archive. The unfixed nature of technological advancement, dependent as it is on in-process decision-making by Internet architects, network designers, platform creators, and policymakers, leaves potential sources of enriching cultural artefacts exposed to constant threat from loss due to server instability, mistranslation, format obsolescence, restrictions on copying, or simple failure to recognize for preservation. As quickly as archival methods are developed, new obstacles arise. While web crawlers have proven indispensable for capturing and preserving sites on the Internet, they are plagued by changes to coding languages used to create websites that render them less effective, challenged by the spread of Flash technology that they cannot currently capture, and purposely undermined by site creators using robots.txt files to block access. There are also practical challenges, including managing the cost of server space, figuring out how to avoid preserving redundant items given the ease of digital copying, and designing user-friendly interfaces to make material accessible once it has been captured.

While digitization efforts enjoy the relative security of being mere duplications of physical objects like parchment and celluloid that have an established institutional tradition of being recognized as endangered and protected by a working set of preservation practices, born-digital materials are far more imperiled by unformed digital heritage policy. The former may expand democratic access to archival resources by effectively removing them from their physical context, but the latter are quickly become the sole record of how individuals, companies, and governments communicate. In addition to the most obvious traditional resources that warrant archiving in their new digital formats, such as news footage, journalism products, television shows, and popular music, there is a vast world of new cultural artefacts being generated through the interactive and participatory culture of Media 2.0 in the form of original creative works, recorded testimonies, personal photography, comments, reviews, and more that are being shared by individual users with wider audiences than ever conceivable even a decade ago.

Though capturing, preserving, and providing access to these new categories of artefacts remains paramount, it is also important to understand the theoretical underpinnings of these projects. The advent of ICT and the networked society imagined by Manuel Castells have forced scholars of media and memory to adjust their ideation of how digital communication and mnemonic practice shape

contemporary cultures. Growing expectations among users that their experiences with media today will prominently feature interactive elements and opportunities to contribute to a larger public discourse are just some of the changes that underwrite the transition to what Andrew Hoskins calls the “new memory ecology.” As a result of the shift to a post-broadcast understanding of how the past is mediated, which defies the simplistic delineation between producers and audiences, he argues that there is now “a greater mixing of the personal and the public, and a routine meshing of the witnesses to events (including those seen as perpetrators and victims) and archives comprising data of these and other events deemed connected” . As media and memory are intertwined, the flexible and evolving nature of digital communication is mirrored in the recent scholarly reimagining of cultural memory as similarly unfixed. Media memory is now best conceived as a process of negotiation between various competing narratives, both in the immediate aftermath of significant events and over time as they are constantly revisited and re-evaluated through the lens of newly possible unbroken access to archives online. Moving forward, it is likely that the same challenge to the traditional authority of media producers of the broadcast age posed by the advent of digital technologies will be paralleled in institutional archival practice thanks to the ability of those same technologies to enable non-professionals to curate some aspects of cultural production, whether intentionally or unintentionally.

While Jack Goldsmith and Tim Wu have used the example of the Principality of Sealand to illustrate that the Internet is not the Wild West free-for-all that some early adopters dreamed about, where nation-states watch helplessly without the ability to regulate transnational commerce originating outside their borders, the capacity of ICT to facilitate the flow of information beyond the traditional boundaries of physical proximity that once limited human interaction has also reshaped how we must think about cultural memory. Anna Reading termed this new digitalized mnemonic terrain the “global memory field,” a place where competing narratives of the past move across physical borders, negotiate around obstacles (like localized censorship), travel faster, resonate more often, flow across mediums more freely, and evolve differently over time than they did in the pre-ICT world. In effect, the cultivation of cultural memory now takes place in the context of digital networks that press remembrances tied to personal experience, as well as those anchored in local, regional, national, and international identities, against one another in a way that causes the echo of the others to gradually reshape each one over time. As an example, consider the difference between the Japanese bombing of Pearl Harbor in 1941 and the terrorist attacks of September 11, 2001. When the earlier event occurred, news of it certainly travelled far and wide, but there was little opportunity for individual citizens to share their reactions with others living across the country, let alone abroad, unless they knew someone personally that they could write to or call (and could afford the long-distance charges).

In contrast, the events of September 11th occurred at a time when those interactions were becoming increasingly common. While the swirl of patriotism around the narrative of “a nation under attack” went largely uncontested for Americans at home during World War II, a similar reaction sixty years later has been tempered by exposure to conflicting interpretations of events coming from those living outside the U.S. A brilliant illustration of the potential impact of the global memory field can be seen in a digital exhibit titled *Where Were You on Sept. 11, 2001?*, an interactive map housed on *The New York Times* website . Users were encouraged to answer the title question in their comments and select a corresponding emotion. Each response was then illustrated as a single dot on a world map. Though responses were most concentrated within the U.S., there is significant comment activity across the rest of North America, Latin America, and Europe, as well as in India. The digital “voices” reflect a full spectrum of possible reactions, including displeasure with the American government’s response, identification with the tragedy

despite not being involved, and personal remembrances, that defy their location on the map. The global impact of 9/11 and the desire of individuals all over the world to participate in the memorialization process is similarly evident in online archives that have collected materials specific to those events, including a collection of digital art at 911digitalarchive.org submitted by users. One example features an assemblage of a personal poem layered over a photo of the New York skyline before the World Trade Center was attacked created by “an Englishman in Saudi Arabia” who wished to pay tribute to those who lost their lives . Another, submitted by a Canadian woman, combines American symbols with the maple leaf icon in tribute . These shifts in user expectations, traditional archival authority, and the scope of mnemonic discourse set the stage for examining the particular ecology of unintentional archives and troubleshooting the consequences of the privatization of communicative cultural heritage.

3. Communication, Inc.: Corporate Spaces and Cultural Production

Recent technological developments have resulted in the privatization of large swaths of digital communication through corporate-owned content-sharing platforms, social-networking sites, and operating system-specific mobile applications. This shift in how much of communication practice is organized and overseen necessitates careful consideration of the potential benefits and concerns brought to light by these new tools. On one hand, innovation has expanded the scope of information exchange and mnemonic discourse beyond the traditional borders of the nation-state as the expense and time required to transmit materials—whether documents, voices, images, or videos—from one corner of the world to the other have become nil. In contrast, the reconfiguration of communication, encompassing everything that was once categorized across a dichotomy of public and private, under the umbrella of transnational corporations has given them a unique power to control the singular discourse and, as an extension, cultural memory. In order to understand the impact of governmental policy on the business practices of content-sharing platforms and the consequences for unintentional archives, a working definition of this category of digital communication tools is necessary.

Corporate content-sharing platforms come in many shapes and sizes. Some focus on particular categories of content. Some emphasize integrated social-networking tools more than others. Some cater to niche categories of users. Unlike peer-to-peer services, these platforms are designed with the goal of aggregating material for temporary access rather than the permanent transfer of files. While the download of files is possible, content-sharing platforms emphasize viewing over caching. For videos, this is accomplished through Flash-enabled streaming of content. For photos, this means browsing through compressed thumbnails rather than full-size raw image files. Access to the original files is often limited, either by denial of access instigated by the user who uploaded it or by site designs that omit the necessary tools to strip out the content. Unlike social-networking sites like Facebook, which also allows users to post images and videos to their profiles, content-sharing platforms spotlight the material more than the personality of the user posting it. While some platforms do allow users to craft an identity on the site and provide contextualizing information, viewing it typically requires clicking through to a secondary page and is certainly not required. For the purposes of this study, I will focus on YouTube, the popular U.S.-based platform that aggregates user-generated video content alongside sponsored content and channels. YouTube has become the default option for millions of users worldwide because the ease of use and the size of its built-in audience, making it ideal for examining how corporate ownership affects the archive of digital cultural production. This does not mean that similar collections of user-generated content on less prominent content-sharing platforms should be overlooked by preservation efforts. This is especially true

given the shifting popularity of sites like these, a fact which only magnifies the need to collect these artefacts before they disappear.

Since its launch in 2005, YouTube has provided seemingly unlimited bandwidth for uploading and storing videos to any user willing to invest a few minutes of their time to complete a brief account creation process and check a box promising that they had read and agreed to the site's terms of service. The price of participation for those who just want to view the videos housed on the site is even lower, requiring no account or login process. Like many Internet startups, YouTube began without a clear plan for monetizing the service provided. As it evolved, the project developed a financial model driven by advertising and corporate sponsorship before ultimately being purchased, a little over a year after it appeared, by the most dominant transnational provider of Internet-based services today, Google. From day one, the incredible growth of the video-sharing site and its monopoly over would-be competitors has been driven by offering the service at no visible cost to users—though the aphorism about nothing being free holds true in terms of the quantity of personal data collected from site visitors in exchange for their use of the site—and developing a sleek user interface that makes posting and finding videos second-nature for all but the most technologically disadvantaged. The success of the site is astronomical, with over 4 billion videos viewed daily and more than 800 million unique visitors each month. Like its parent company, the scope of YouTube extends well beyond its origins as an American company, with over 70 percent of traffic coming from locations outside the United States and localized sites in approximately 50 languages in 39 countries .

Following the theoretical trajectory marked by practitioner-scholars—from librarian Karen Gracy , who problematized the shifting authority of professional archivists in light of the growing popularity of content-sharing platforms, to archivist Rick Prelinger, who highlighted the absence of social responsibility embodied by these sites when compared to formal analogue archives of cultural heritage—Jean Burgess and Joshua Green have defined YouTube as an “accidental archive” of cultural production. They argue that YouTube is unique in that has collected billions of hours of valuable material, everything from old advertisements and media clips to original creative works, despite the fact that this content is aggregated on the site for fleeting use with little recognition of its importance as a resource for future scholars, archivists, librarians, and the general public. Videos are accessible primarily through search algorithms rather than explicitly curated and are posted with no promise of lasting preservation. While Burgess and Green touch on all of the important features of the YouTube phenomenon, the collective assembly of cultural material on this content-sharing platform hardly seems accidental. The clustering of communication practice around technologies that allow individuals to broadcast themselves to the broadest audience at the lowest cost using the simplest interface is not a new or surprising trend. The repository of user-generated content on YouTube may be better identified as an “unintentional archive,” which emphasizes the absence of curation and attention to preservation that characterizes traditional archives and further delineates this material from intentional attempts to widen access to archival material through digitization being conducted by traditional cultural heritage institutions on these same platforms. While the U.S. Library of Congress and the Smithsonian Institute both currently maintain channels on YouTube that replicate a small percentage of content from their physical archives among other public engagement efforts, the focus of the unintentional archive is exclusively user-generated born-digital cultural production.

Beyond the technological threats to preservation shared with online archives intentionally created and maintained by traditional cultural heritage organizations, the location of unintentional archives within the ecology of privately-owned content-sharing platforms like YouTube makes them subject to additional

imperilment as a result of market economic factors. These sites, housed under the corporate umbrella of publicly-traded transnational media companies, are responsible first and foremost to their investors. This means turning a profit, an outcome often at odds with the expectations and practices of site users. Burgess and Green provide one example of how the pursuit of profits can conflict with the organic ways that users have appropriated the tools provided for their own purposes, describing initial user backlash against the addition of channels sponsored by legacy media institutions . While this commercial element has become the new norm on the platform, the incident is a useful reminder that the way that sites like YouTube are used today is not necessarily how they were originally envisioned by their creators, nor is their continued use in that manner guaranteed. This is particularly true of their role as sites of unintentional archival practice.

To fully understand what is at stake in failing to acknowledge the value of unintentional archives and properly develop strategies for counterbalancing the privatization of digital heritage, imagine what would happen YouTube was shuttered tomorrow. Where would the content go? Who would be able to access it? The potential danger for these digital artefacts extends beyond the worst-case scenario of a business failure. What if the corporation simply decided to shift focus because the current model was no longer generating sufficient profits? What if the site's functionality remained largely the same but the company rolled out a radically different TOS agreement? What if your local political leaders suddenly decided to pressure the service provider or its parent company to adhere to strict policies for censoring certain content or be forced to quit doing business in that country, as already happens under restrictive regimes like those in China, Pakistan, and India? The policies dictating the use of sites like YouTube make no promises about the future. This uncertainty is further aggravated by the location of privately-owned media companies beyond the reach of nation-state governance of important issues like free expression.

4. Where the First Amendment Ends: From the Constitution to Terms of Service

When culture is constructed and maintained through privately-controlled domains, the Constitutionally-protected rights of individuals, like the freedoms of speech, religion, and assembly enjoyed in the United States, are no longer guaranteed. As Jack Balkin has argued, the First Amendment, which ensures those rights for all Americans, no longer has as much relevance in the information age. He purports that this shift is not because the free expression of ideas and the unencumbered exchange of information have become somehow less important, as the right of individuals to contribute to democratic culture remains an essential liberty . In fact, free expression online has traditionally been encouraged by Section 230©(1) of the Communications Decency Act of 1996, a provision which releases all manner of Internet content providers from liability for the speech of individuals posting to their platforms, including that which might be categorized as hate speech, defamation, or libel. While this theoretically minimizes the potential motivation for these sites to actively police and censor users, this does not reflect common practice. They universally reserve the right to remove content from their sites whenever they see fit and terminate your account without explanation. With few exceptions, the U.S. federal government is not in the business of forcing companies to serve every potential customer, so content-sharing platforms can set whatever policies they want and the only recourse for users who disapprove is to stop utilizing that service. The law of the nation-state is effectively replaced by the law of the corporation, with the latter being far more absolute. While state and federal statutes governing conduct in the United States are subject to the promise of due process and open to some possibility of appeal through the judicial system, users of

corporate content-sharing platforms are subjected to all-or-nothing scenario. Whether they realize it or not, users automatically subject themselves to all stated policies as soon as they visit one of these sites and companies are not shy about reminding users that they are free to leave if they are not satisfied with the terms of use.

While First Amendment protections for individuals are undercut when communication moves to content-sharing platforms, the corporations behind these sites enjoy one very important right granted by the federal government through the Digital Millennium Copyright Act (DMCA) of 1998: immunity from prosecution for copyright infringement related to user-posted content if they comply with notice-and-takedown procedures that require the removal of questionable material immediately upon request from copyright holders. This additional protection is necessary because Section 230 of the CDA does not apply in cases of intellectual property violation. Combined with ever-expanding protections for intellectual property that disproportionately favor copyright holders, especially legacy media conglomerates, the DMCA is the largest factor shaping how U.S.-based content-sharing platforms do business. As profit-seeking enterprises, sites like YouTube—which traffics heavily in popular culture artefacts that are inescapably protected by copyright—are prime targets for infringement litigation. Unlike individual users posting videos to the site, YouTube’s parent company, Google, actually has the deep pockets required to pay out on the substantial fines and court costs associated with conviction for intellectual property violations. Not wanting to bother with litigation or risk potential financial penalty, these platforms err on the side of caution when it comes to responding to accusations of infringement. Wendy Seltzer has argued that the net effect of this “safe harbor” provision in the DMCA is a further chilling of free expression for users, as content ranging from campaign videos to home movies to transformative creative works has been removed in response to notice-and-takedown orders with very limited opportunity for individuals posting such content to argue against it. Though it is the only Constitutional privilege Google and Yahoo! uphold religiously, the enforcement of the intellectual property rights of content creators still hinges on the policies of the individual company created in the context of the law rather than direct government legislation. The language governing how each platform manages notice-and-takedown orders is established by its TOS agreement and, as very few infringement cases ever make it to trial thanks to the DMCA, there is little opportunity for intervention on behalf of users whose content was wrongly censored. This is only magnified by the transnational presence of YouTube and the fact that it rarely adjusts its business model as it expands into other countries. Given the concerns for free expression, this casual modeling of U.S. policy norms in countries all over the world through YouTube’s practices, combined with the effects of formal international copyright escalation like the Anti-Counterfeiting Trade Agreement (ACTA), is dubious at best.

With its dense legal prose, the YouTube TOS agreement follows the same basic outline as other content-sharing platforms, sketching out a variety of responsibilities and restrictions for users alongside important exemptions for the company. Though they have occasionally launched widespread campaigns to notify users of significant policy revisions and required individuals to acknowledge their compliance by clicking “accept,” tacit agreement to the company’s terms does not require registering for an account. If you do not agree with any aspect of the terms, are under 13 years old, or are otherwise incapable of issuing consent, you are instructed not to visit the site at all. Much of the language, especially prominent sections of all-caps text, is dedicated to releasing the corporation from any and all liability for any personal injury, property damage, defamation, criminal prosecution, loss of content or service, illegal appropriation of your content by others, or exposure to false or obscene content that may arise from use of the platform. If you have an account and it is misused without your permission, you are liable but

YouTube is not. Other sections focus on what users are not allowed to do, including distribute, modify, access, or commercialize any aspect of the service or its content unless through YouTube approved channels and with the company's permission. User accounts may be cancelled at any time at the discretion of the company without justification and content uploaded to the site is never returned. A special subsection, called "Community Guidelines," celebrates individual free expression half-heartedly while laying out types of content that will be censored by the site, including videos featuring sexually explicit content, graphic violence, gross or shocking material, or other "bad stuff like animal abuse, drug abuse, under-age drinking and smoking, or bomb making" . Harassment, invasion of privacy, and spamming are also prohibited at the discretion of the site, but no concrete information is provided about how these issues are handled other than by flagging videos for review by YouTube staff.

Regarding intellectual property, the TOS agreement states that the act of posting a video assumes that the user has addressed all potential copyright concerns and has the rights to do so. If this is not the case, liability for intellectual property violation is placed solely on the individual users. Uploading content to the site automatically grants YouTube "a worldwide, non-exclusive, royalty-free, sublicenseable and transferable license to use, reproduce, distribute, prepare derivative works of, display, and perform the Content," though the user retains their ownership rights. The remaining value of those ownership rights, given the breadth of the license YouTube grants itself to your content as soon as you post it, is not explained. The license on videos expires within the vague window of "a commercially reasonable amount of time" after you delete them, however the company retains perpetual rights to use of all comments made on the site as it sees fit. Even when the video license expires, the company reserves the right to keep copies of deleted content even though it may not display them, rendering the promise of users being able to maintain ownership rights over their original content seem even less relevant. Finally, the processes for notification and counter-notification related to copyright infringing material in compliance with the DMCA are included as part of the terms of service, though little contextual information for how YouTube will actually process complaints is provided.

There are a number of provisions of the YouTube's TOS agreement that are particularly relevant to the unintentional archives of cultural production aggregated on the platform and potential efforts to capture and preserve that content. With the exception of public search engines, automated web crawlers are prohibited. Where permissible, they are only allowed to index content without caching it. Furthermore, despite numerous third-party web browser applications that provide this function, downloading content is forbidden unless facilitated through a service provided by YouTube. The language in the TOS agreement is sufficiently vague that this would seem to apply to users who wanted to regain access to their own content, in the case of their copy becoming corrupted or lost, as well. Users are not allowed to "copy, reproduce, distribute, transmit, broadcast, display, sell, license, or otherwise exploit any Content for any other purposes without the prior written consent of YouTube or the respective licensors of the Content," which would put the burden on non-profit online archivists, such as the Internet Archive, and traditional cultural heritage institutions to ask for permission and gain agreement before carrying out any outside preservation program. Finally, the TOS agreement announces the company as an exclusively American, subject only to the federal laws governing the U.S. and the state and local laws of California, where its physical operations are located. Users in other countries are thereby required to interpret their use of the site through the additional frame of the laws where they live on their own. While it covers a significant amount of legal ground, the overall effect of the TOS agreement, unsurprisingly, is to limit any potential use of the site or its content in ways not approved by YouTube, to release the company from all liability by placing it on individual users, and to grant the company exclusive license to

manage the site however they choose. The result is the severe limiting of both the potential for unencumbered self-expression on a popular digital communication platform and the possibility of adequately documenting the unintentional archive of cultural production housed there.

5. Mapping the Future: Three Emergent Models of Born-Digital Cultural Heritage Preservation

The privatization of digital heritage as communication practice shifts to corporate content-sharing platforms is a problem without a clear solution. The first step—raising awareness about the existence of unintentional archives of important cultural artefacts on the sites—is always the easiest. Through papers like this one and interdisciplinary collaboration between media scholars, information technology experts, policymakers, and non-profit advocates, it is possible to spread the message of the value of archiving born-digital cultural production and the urgency with which we must develop strategies to preserve it broadly. Unfortunately, awareness is not enough. Implementing a plan of action is complicated by the emergence of the new memory ecology and the globalization of mnemonic practice as a result of ICT. The free flow of information across physical borders has transformed the issue of preserving heritage from one of local concern, easily managed within traditional nation-state borders, to one that concerns us all. To date, three models for managing the preservation of born-digital content have emerged. One model depends on the benevolence of corporate content-sharing platforms to recognize the import of what they have collected and take steps to share and preserve it as part of the historical record. The second prototype looks to government policymakers to repair the state of copyright in their own countries in order to encourage access to cultural production and then work together on an international scale to develop archival practices that reflect the global nature of communicative culture today. The final scenario involves non-profit preservation organizations driving change by designing tools to gather content from these private sites, charging headfirst into legal gray areas or even by openly flouting existing intellectual property laws, and collaborating with traditional cultural heritage institutions. Each approach certainly has its benefits and flaws and exploring them will be helpful in guiding best practices for the future.

As I have stated previously, the profit-driven motives of corporations often run counter to public interest in the greater good. This is why the self-determined laws governing activity on these sites never account for preservation or make promises for the future. Part of what makes some enterprises more successful than others are flexibility and adaptability in the face of change, whether technological, economic, political, or social. Laying out a plan for ensuring access to their products and services decades or even hundreds of years from now is neither a priority nor particularly beneficial to the bottom line. Still, one platform has responded to its significance as dominant mode of communication by exploring options for digital preservation. In 2010, Twitter announced that it would partner with the Library of Congress to archive every public tweet since the site launched four years earlier . Made possible by a legal document called a “Gift Agreement,” Twitter donated a complete copy of its current archive and all tweets going forward, including whatever copyright license the platform had claimed on the user-generated content, to the Library with the stipulation that any of it could be made available to the public after a 6-month window from the date of original posting had passed. Managing the massive quantity of data and creating a user interface to allow researchers to access it in the future is an undertaking that is still in progress, the burden of which has been placed entirely on the Library of Congress . Among the more obvious downsides of this type of arrangement is the fact that the material, while duplicated on the servers of a traditional cultural heritage institution with a long track record of archival responsibility and

experience managing digital content, none of it is actually accessible to the public for research or review two years later. Nagging concerns also persist over whether the complex copyright issues of who owns tweets have been fully resolved and whether the individuals who originally generated this enormous backlog of data should have any say in this repurposing of their content. As Twitter is a transnational platform like YouTube, there is the serious question of whether negotiations between an American company and a U.S. cultural heritage institution should determine the fate of content produced by users across the globe. Most importantly, this model depends heavily on the self-aware public relations-driven benevolence of corporations, which are notorious fickle when it comes to matters of public interest, to take action. As Twitter has recently decided to monetize the same content archive by packaging it for data-mining companies who want to sell the information they can glean from it to marketers, it seems unlikely that this is a sound means of guaranteeing the security of all cultural production housed on content-sharing platforms for future generations .

The second scenario is twofold, requiring governments around the world to enact legislation that would lower some of the barriers to preservation posed by stringent copyright protections and then actively invest in the development of collaborative digital archival projects. This appears to be the least promising avenue for protecting unintentional archives. Copyright escalation has become a global issue, with the restrictive legislation of the U.S. being mirrored in countries far and wide, whether through internal revisions to nation-state policy as in Canada or through arm-twisting trade agreements like ACTA that encourage compliance among smaller nations using the carrot of beneficial commerce exchange. As the U.S. model increasingly becomes the default standard for copyright worldwide, it is important to note that there is far from universal support for the current legislation even in America. Non-profit advocacy group Public Knowledge has widely promoted its list of reforms that U.S. copyright laws need desperately to preserve the marketplace of ideas and facilitate access to cultural production, including shortening terms of protection, ending DMCA abuses, removing the unilateral ban on DRM-cracking, and increasing support for “fair use” in keeping with the knowledge-building spirit of the original framework for copyright found in the Constitution . Unfortunately, the momentum of copyright legislation is moving strongly in the opposite direction. Inter-governmental coordination of collaborative digital archiving initiatives has been somewhat more successful, with projects like the World Digital Library beginning to take shape at the behest of the Library of Congress and UNESCO. While the potential value of this resource in terms of global heritage is quite impressive, the focus so far has been on aggregating digitized reproductions of physical artefacts so the overall impact on born-digital archival practice and unintentional archives specifically is still unclear.

The third potential model for capturing and preserving born-digital cultural production on corporate content-sharing model uses non-profit organizations like Internet Archive in the U.S. and the Internet Memory Foundation, based in Europe, to spearhead their own archival projects, develop tools for collecting and curating content, and partner with traditional cultural heritage institutions. Founded in 1996 by web archiving advocate Brewster Kahle, Internet Archive is active on all three of these fronts and helped to establish the International Internet Preservation Consortium (IIPC), a coalition of almost 40 national libraries worldwide that collectively determines best practices such as standardized formats for website caching and divides up responsibility for capturing the whole of the web . Internet Archive is also widely recognized for its Robin Hood-like efforts to capture as much content as possible using web crawlers that copy sites without seeking preordained permission from site owners. While site administrators are free to use robots.txt files to block their pages from being crawled and/or to request that Internet Archive not capture their content, these stopgap measures do not always work. Working on the

general premise that what they are doing is part of an exigent cultural imperative to preserve communication taking place online today for posterity, they have yet to encounter many legal challenges, although use of their archive of webpages as evidence in business litigation has raised some initial concerns among corporate lawyers about the unapproved duplication of their client's sites. While the law has yet to catch up with the realities of digital cultural production and the need to preserve that material, the archival projects that Internet Archive has already implemented are impressive. Their "Understanding 9/11" archive of news footage from the week of the attacks, made possible by the organization's decision to begin recording digital broadcasts from 20 stations around the world on 24/7 basis in 2000, is an incredible research tool for media scholars. Their "Wayback Machine," a simple interface for accessing archived websites, allows users to pull up every version of websites that have been captured by their web crawlers. This includes more than 4,600 versions of YouTube captured between 2005 and 2011. While this is obviously an incomplete record of the content-sharing platform given how often new videos are uploaded and how many of the links on older versions of the site are now broken, it does hint at the achievable outcomes when motivated, innovative non-profits team with traditional cultural heritage institutions to accelerate the learning curve for digital archival practice.

Of these three emergent models, the latter seems most appropriate for ensuring the preservation of born-digital cultural production found in unintentional archives on corporate content-sharing platforms. The unique position of non-profits at once inside and outside the traditional heritage preservation system allows them the access and authority of working with legacy institutions as well as the flexibility to forge into uncharted legal territory in a way that government-affiliated libraries and archives cannot. As technologists, these passionate advocates have had great success convincing librarians and archivists in formal settings of the value of capturing the digital cultural record. This would naturally carry over into raising awareness about the particular sensitivity of unintentional archives and leading the charge into private domains where important public exchange is currently taking place. While the corporate benevolence and governmental reform scenarios promise uncertain results, non-profit partnership has already led to important improvements in coordinating born-digital preservation on the global scale required by the new memory ecology. It has also delivered concrete resources which enable researchers to work with these materials immediately and serve as a model for what can be accomplished in order to encourage traditional cultural heritage institutions around the world to take decisive action. The legal uncertainties surrounding some of their practices is a concern, however the fact that preservation techniques are constantly playing catch-up to rapid innovation in the field of digital communication and the reality that legislative reform moves even more slowly may justify this kind of action as the only way to prevent the loss of this material until they evolve. Non-profit partnership has also been essential in forging collaborative ties between distant traditional cultural heritage institutions, perhaps the most essential factor that will determine the success of born-digital archiving initiatives of the future. Coordination and standardization of these new archival practices across the global memory field is the only means of ensuring that the digital cultural record will be preserved in a way that accurately reflects its significance to contemporary culture.

References

- Anonymous. "International: Born Digital; Archiving the Web." *The Economist* (October 23, 2010): 73.
- . "Internet Archive Raises Copyright Concerns." [In English] *Inside Counsel*, no. 19306393 (2006): n/a.
- Auntmannys. "Virginia Tech Memorial (Video)." <http://www.youtube.com/watch?v=d6L42Ijzfc>.
- Balkin, Jack M. "Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society." *New York University Law Review* 79, no. 1 (2004): 1-55.
- . "The Future of Free Expression in a Digital Age," *Pepperdine Law Review* 36(2008-2009): 427-444.
- Burgess, Jean, and Joshua Green. "Youtube's Cultural Politics." In *Youtube: Online Video and Participatory Culture*, 75-99. Cambridge, MA: Polity, 2009.
- Carey, James. "A Cultural Approach to Communication." In *McQuail's Reader in Mass Communication Theory*, edited by Denis McQuail, 36-45. London, UK; Thousand Oaks, CA: Sage Publications, 2002.
- Castells, M. *The Rise of the Network Society*. Wiley-Blackwell, 2009.
- Clark, Jessica, and Patricia Aufderheide. "Public Media 2.0: Dynamic, Engaged Publics." Center for Social Media. <http://www.centerforsocialmedia.org/future-public-media/documents/white-papers/public-media-20-dynamic-engaged-publics>.
- Featherstone, Mike. "Archive." *Theory, Culture & Society* 23, no. 2-3 (May 1, 2006): 591-96.
- Goldsmith, Jack L., and Tim Wu. *Who Controls the Internet?: Illusions of a Borderless World*. New York: Oxford University Press, 2006.
- Gracy, Karen F. "Moving Image Preservation and Cultural Capital." [In English] *Library Trends* 56, no. 1 (2007): 183-97.
- Hoskins, Andrew. "Anachronisms of Media, Anachronisms of Memory: From Collective Memory to a New Memory Ecology." In *On Media Memory: Collective Memory in a New Media Age*, Palgrave Macmillan Memory Studies, edited by Motti Neiger, Oren Meyers, and Eyal Zandberg, 278-88. New York, NY: Palgrave Macmillan, 2011.
- Kahle, Brewster. "The 9/11 Television News Archive." Video Recording of Presentation, Learning from Recorded Memory: 9/11 Tv News Archive Conference, New York, NY, August 24, 2011. <http://archive.org/details/911conferenceBrewsterKahle>.
- Lasar, Matthew. "Why Wait? Six Ways That Congress Could Fix Copyright, Now." 2012. <http://arstechnica.com/tech-policy/news/2012/02/why-wait-six-ways-that-congress-could-fix-copyright-now.ars>.
- Liddle, Janice. "Freedom's Flame Will Forever Shine Brightly." http://static.911digitalarchive.org/REPOSITORY/IMAGES/DIGITAL_ART/2062.jpeg.
- McMillan, Graeme. "Tweet Eternal: Pros and Cons of the Library of Congress Twitter Archive." 2011. <http://techland.time.com/2011/12/08/tweet-eternal-pros-and-cons-of-the-library-of-congress-twitter-archive/>.
- Prelinger, Rick. "Archives and Access in the 21st Century." *Cinema Journal* 46, no. 3 (2007): 114-18.
- Raymond, Matt. "How Tweet It Is!: Library Acquires Entire Twitter Archive." 2010. <http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/>.

- Reading, Anna. "Memory and Digital Media: Six Dynamics of the Global Memory Field." In *On Media Memory: Collective Memory in a New Media Age*, Palgrave Macmillan Memory Studies, edited by Motti Neiger, Oren Meyers and Eyal Zandberg, 214-52. New York, NY: Palgrave Macmillan, 2011.
- Seltzer, Wendy. "Free Speech Unmoored in Copyright Safe Harbor: Chilling Effects of the Dmca on the First Amendment." *Harvard Journal of Law & Technology* 24 (2010): 171.
- Taylor, Graham. "The Twins." 2012.
http://911digitalarchive.org/REPOSITORY/IMAGES/DIGITAL_ART/1148.pjpeg.
- Wasserman, Todd. "Twitter Is Selling Your Old Tweets." 2012. <http://mashable.com/2012/02/28/twitter-is-selling-old-tweets/>.
- Watters, Audrey. "How the Library of Congress Is Building the Twitter Archive: Checking in on the Library of Congress' Twitter Archive, One Year Later." 2011.
<http://radar.oreilly.com/2011/06/library-of-congress-twitter-archive.html>.
- Woodruff, Andy, David Heyman, Ben Sheesley, Shan Carter, and Jeremy White. "Where Were You on Sept. 11, 2001? (Interactive Map)." 2011.
<http://www.nytimes.com/interactive/2011/09/08/us/sept-11-reckoning/where-were-you-september-11-map.html>.
- YouTube. "Community Guidelines." http://www.youtube.com/t/community_guidelines.
- . "Terms of Service." <http://www.youtube.com/t/terms>.
- . "Youtube Statistics." http://www.youtube.com/t/press_statistics.

Context 2.0

User Attitudes to the Reliability of Archival Context on the Web

Heather Ryckman

Abstract

The representation of context is the foundation of archival practice. With the phenomenal growth and popularity of Web 2.0 archival institutions are increasingly making their collections available online through wikis and photo-sharing sites. However, the structure of Web 2.0 sites and the inclusion of user contributed content have changed both the content and presentation of contextual information. This paper examines attitudes to the reliability of archival description online, including the role of user contributed content to the understanding of context, and draws conclusions on the function of archival contextual information online and its impact on users' understanding of context.

Author

Heather Ryckman has been working in the heritage field since 1995. In addition, to working as the Assistant Archivist for the Bank of Canada, Heather has held curatorial and collections management roles for such institutions as the Markham Museum, the Wallaceburg Museum and the Burlington Museums. She has also served as instructor for the Archives Association of Ontario's exhibit design and volunteer management seminars. She holds a BA in Ontario History and Anthropology, and an MA in Public History with areas of specialty in Museology and Ontario History. Heather has been employed as the corporate Archivist for The Co-operators Group of companies since May 2007.

1. Introduction

The documentation of context is the foundation of archival arrangement and descriptive practice, and is acknowledged as being integral to maintaining the authenticity and evidentiary value of archival records. Despite varying arrangement and descriptive processes worldwide for recording and presenting archival information, national and international standards acknowledge the primary importance of context in the understanding and interpretation of archival records.¹ The modern electronic recordkeeping environment has necessitated an expanded, more comprehensive, and inclusive view of both the nature of context and those who can and should interpret it. In addition to the increasing role of archivists as sources of knowledge, there is the recognition that archival users collectively may have more knowledge about archival records than an archivist,² and that there are many possible and constantly changing interpretations and provenances that can be ascribed to every archival record.

The online electronic environment has provided archivists with the opportunity to make their institutions and collections more visible and accessible to wider audiences. Yet in the race to make collections accessible, many institutions have changed the way in which they present and structure archival

¹ Some examples of standards which recognize the importance of context in archival description include: RAD (Rules for Archival Description), ISAD-G (International Standard for Archival Description, General), ISAAR-CPF (International Standard Archival Authority Record for Corporate Bodies, Persons and Families), DACS (Describing Archives: A Content Standard), and the Commonwealth Records Series system.

² Isto Huvila, "Participatory archive: towards decentralised curation, radical user orientation, and broader contextualisation of records management," *Archival Science* 8 (2008): 15-36, doi:10.1007/s10502-008-9071-0.

descriptions online. Web 2.0 has been heralded as a great opportunity for archivists to increase their visibility and serve an increased number of users. In reference to Web 2.0 sites such as Flickr, “the implications of this phenomenon are significant in that patrons will have increased information and description of a resource, an enhanced ability to find resources through search and browsing, and use of the resource as a connector to additional materials on the web.”³ Additionally, the openness and inclusivity of Web 2.0 is seen by many scholars as a means of permitting multiple voices and interpretations to enter archival descriptions, and provide a more comprehensive understanding of the nature and context of archival records. Yet as the archival community’s involvement in Web 2.0 initiatives grows, the completeness and accuracy of the contextual information become increasingly important.

Through an analysis and survey of case study websites, this paper will examine the impact of Web 2.0 practices on the presentation and ultimate perceived reliability of contextual information online.

2. Literature Review

Knowledge and understanding of context⁴ have long been recognized by the archival community as essential to the complete understanding of a record’s structure and content. As early as the late 1800s, historical investigation was used to provide an understanding of a record’s meaning.⁵ This meaning imbues records with the reliability, authenticity and usability required to be trusted and understood sources of evidence. In order for users to assess the validity of a record, users must be able to access the contextual information pertaining to the creation, custody and use of the record.

The nature of digital records has challenged archivists’ traditional understanding of the extent of context. Digital objects, unlike physical ones, lack the history and context of interaction and association that provides valuable information to the user and interpreter of the object.⁶ The ability of digital records to be imperceptibly changed and easily integrated into many different contextual environments affects both the nature of the context that needs to be documented, and the manner in which it is recorded. Additionally, the lack of transparency associated with the creation, alteration and use of digital records has resulted in a recordkeeping environment that is increasingly reliant on both context and trust as measures of the authenticity of a record. In assessing the integrity of a digital object, researchers must

³ Michael Zarro and Robert Allen, “User-Contributed Descriptive Metadata for Libraries and Cultural Institutions,” in *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science*, ed. Mounia Lalmas et al. (Springer Berlin / Heidelberg, 2010), 46.

⁴ The terms ‘context’ and ‘provenance’ will be used interchangeably in this paper. Despite the significance of the principle of provenance to all aspects of archival practice, there is no consensus as to the extent of the term and the best way to apply it. See Shelley Sweeney, “The Ambiguous Origins of the Archival Principle of “Provenance,”” *Libraries & The Cultural Record* 43(2008): 193-213. While the traditional notion of provenance was limited to the origins of the record, many modern archival scholars, including Terry Cook, “Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era,” *Archives and Manuscripts* 22(1994): 300-328, and Laura Millar, “The Death of the Fonds and the Resurrection of Provenance: Archival Context in Space and Time,” *Archivaria* 53(Spring 2002): 2-15, have argued for an expanded and more elastic view of provenance, and appear to use the terms “context” and “provenance” interchangeably. In addition to highlighting the lack of professional consensus regarding the meanings of both context and provenance, this suggests that concern over distinguishing between the two terms in the literature may be unnecessary.

⁵ Arnaldo D’Addario, “The Development of Archival Science and Its Present Trends,” in *Archival Science on the Threshold of the Year 2000*, ed. Oddo Bucci (Macerata: University of Macerata, 1992), 183.

⁶ Alan Wexelblat and Pattie Maes, “Footprints: History-Rich Tools for Information Foraging” (paper presented at the SIGCHI Conference on Human Factors in Computing Systems, 1999): 270-277.

“examine the provenance of the object...and the extent to which [they] trust and believe this documentation as well as the extent to which [they] trust the custodians themselves.”⁷

Initially, the archival presence online was characterized by the online finding aid. It is not only the finding aid, though, that is now being made available on the Internet. Archival and non-archival repositories alike are presenting digital representations of their holdings online. Some scholars suggest that the Internet provides a venue to allow records to be interpreted in multiple ways, which serve to “reveal complex and shifting meanings without abandoning a foundation from which such meanings can be gleaned,”⁸ thereby recognizing the complexity and constant meaning-making associated with all records. Other scholars note that Internet research relies on ad-hoc classification and associations, which shift based on the search parameters.⁹ This may result in search results and associations that are determined more by the search terms than by a record’s context.

The growth of Web 2.0 has continued to alter the role and priorities of archives online and resulted in a change in the content and presentation of context. “Using social media tools, archivists even invite user contributions and participation in describing, commenting, and re-using collections, creating so-called collaborative archives.”¹⁰ Collaborative archives are seen as a means of including the voices and interpretations of those unrepresented or under-represented within the archival record, thereby creating a more holistic view of the record and a shared identity and sense of community. Current recordkeeping and archival descriptive systems do not support multiple, simultaneous provenance or co-creatorship involving the exercise of mutual rights and responsibilities in records by all participants in the transactions they document.¹¹ The more openly structured environment of Web 2.0 may be more conducive to supporting and representing simultaneous provenance.

The characteristics and possibilities of Web 2.0, in particular its inclusivity of multiple voices and interpretations, are seen by many archivists as a panacea for the singular authoritative voice of the traditional archival finding aid. MacNeil considers the Web as an,

...ideal vehicle for transcending the artificial limits imposed by current descriptive practices and for exploiting an expanded vision of archival description; one that unseats the privileged status currently accorded to the standards-based finding aid and repositions it as part of a complex network of hyperlinked and interactive documentation relating to the history, appraisal, preservation, use, and interpretation of a body of records over time.¹²

Proponents of collaborative archives argue that user contributed content allows for discussion and discovery, and provides memories and knowledge about records which would otherwise go uncaptured.¹³

⁷ Clifford Lynch, “Authenticity and Integrity in the Digital Environment: An Explanatory Analysis of the Central Role of Trust,” in *Authenticity in a Digital Environment* (Washington: Council on Library and Information Resources, 2000), 34.

⁸ Emily Monks-Leeson, “Archives on the Internet: Representing Contexts and Provenance from Repository to Website,” *The American Archivist* 74(2011): 42.

⁹ Lilly Koltun, “The Promise and Threat of Digital Options in an Archival Age,” *Archivaria* 47(1999): 114-135.

¹⁰ Kate Theimer, “What Is the Meaning of Archives 2.0?,” *The American Archivist* 74(2011): 61-2.

¹¹ F. Upward, S. McKemmish, and B. Reed, “Archivists and Changing Social and Information Spaces: A Continuum Approach to Recordkeeping and Archiving in Online Cultures,” *Archivaria* 72(2011): 230.

¹² Heather MacNeil, “Picking Our Text: Archival Description, Authenticity, and the Archivist as Editor,” *The American Archivist* 68(2005): 278.

¹³ See Andrew Flinn, “‘An Attack on Professionalism and Scholarship’?: Democratising Archives and the Production of Knowledge,” *Ariadne* (2010), <http://www.ariadne.ac.uk/issue62/flinn/>. See also Huvila, “Participatory

This is particularly true of communities which do not share their knowledge with outsiders, including a professional archivist.¹⁴ Yet the increased facility of discovery and discussion of records also threatens to weaken the role of context. The indexing of user contributed comments, which are then searchable by search engines such as Bing and Google and result in researchers being directed to archival resources and related user comments, may pose a problem if the comments the researchers were directed to view are inaccurate. The exporting or copying of the archival records to external, personal collections, or other secondary sites such as Wikipedia, is also problematic, because within these sites the original context of the record, as presented by the archives, may be lost or altered. Even when links between abridged Web 2.0 descriptions to more complete Web 1.0 finding aids are created to provide context to users, research suggests that there was considerably less redirection than expected.¹⁵ These results suggest that providing contextual information only in the Web 1.0 environment and linking to it in the Web 2.0 environment may not be sufficient to give researchers the necessary context.

Folksonomies, often created by 'tagging', are a common method for archives to elicit user contributed content, though there is much debate surrounding the relevance and value of this content. One of the primary values of tagging is that it permits many voices, experiences and interpretations to be reflected in descriptions. As Weinberger notes, "an author is an authority when it comes to what she intended her work to be about, but not about when [sic] it means to others. When it comes to searching, what a work means to the searcher is far more important than the author's intentions."¹⁶ Yet, if understanding an author's intent is essential to the understanding and preservation of context, it has yet to be determined how well folksonomies will be able to represent this context of creation. Bearman and Trant note that "museum collections remain relatively inaccessible even when 'made available' through searchable online databases"¹⁷ due to the professional terminology employed. The use of folksonomies, written in the language of the user, may increase the user's ability to retrieve the required records, but it is unclear whether folksonomies assist the user to understand the records better once retrieved.

User comments and notes are also frequently cited in the literature as a means of allowing multiple voices and interpretations to be heard.

The user searching for an image likely has little or no concept of the context in which the tag was submitted, and therefore may be presented many results in a search that are in fact not relevant in context. In contrast, the comments and notes we studied appear to be meant almost entirely intended to add context to an image.¹⁸

By allowing other perceptions and interpretations to describe context, there is the worry that inaccurate or poor descriptions may result; descriptions which threaten the authenticity and reliability of the records in question. The anonymity of user contributed content "makes it difficult to determine

Archive: Towards Decentralised Curation, Radical User Orientation, and Broader Contextualisation of Records Management."

¹⁴ Flinn, "'An Attack on Professionalism and Scholarship?': Democratising Archives and the Production of Knowledge."

¹⁵ See Martin Kalfatovic et al., "Smithsonian Team Flickr: A Library, Archives, and Museums Collaboration in Web 2.0 Space," *Archival Science* 8(2008): 272-273.

¹⁶ David Weinberger, "Tagging and Why It Matters," Berkman Center for Internet and Society, http://cyber.law.harvard.edu/publications/2005/Tagging_and_Why_It_Matters.

¹⁷ David Bearman and Jennifer Trant, "Social Terminology Enhancement through Vernacular Engagement: Exploring Collaborative Annotation to Encourage Interaction with Museum Collections," *D-Lib Magazine* 11(2005), <http://www.dlib.org/dlib/september05/bearman/09bearman.html>.

¹⁸ Zarro and Allen, "User-Contributed Descriptive Metadata for Libraries and Cultural Institutions," 51.

whether information is biased, since users cannot know the motives for information provision, and the lack of cues about the expertise of contributors similarly inhibits users' capacity to determine the accuracy of information provided."¹⁹ While archival descriptions are by no means objective there is an attempt on the archivist's part to be aware of their own bias. Popular or public interpretations of the records may lack this awareness. Additionally, the archivist likely has access to more information about a record's context than does an online user, who determines his or her tags on an item level, or at best, based upon a small sampling of documents belonging to a much larger whole. If the descriptions presented are wholly inaccurate, or inflammatory, there is little guidance in the literature about how the public evaluates the interpretations of context presented and how these various interpretations impact the overall authenticity of the archival record. At present, the vast majority of contributors to archival sites participate at no more than a cursory level to the archival descriptions, identifying individuals and correcting errors. Few extrapolate or expand upon an existing record description. This paucity of user contributed information may not be enough to represent reliably the multiple voices that the archives is seeking.

3. Research Methodology

To determine the presentation of context online, and attitudes towards this presentation, this paper poses the following research questions:

RQ1: What differences exist in the representation of context between Web 1.0 and Web 2.0 pages?

RQ2: What differences exist between patterns of use in Web 1.0 and Web 2.0 pages?

RQ3: What user attitudes exist towards the representation of context in the case study pages?

This research was carried out within an integrated post-custodial and post-modern theoretical framework as elaborated upon by Terry Cook.²⁰ The recognition of the fluidity and complexity of archival context forms the foundation for the role and importance of Web 2.0 and user contributed content in archival description. There is no common position among archivists regarding the extent of information required to document a record's context properly, or to represent this context online. What can be asserted is that the following types of documentation enable users to make sense of the context of records online:

1. Creator Documentation: information that explains the circumstances surrounding the creation of the record. This can include the name of the record's creator, the function, mandate and administrative structure of the record creating body, and the social, political and economic environment in which the record was created
2. Relationship Documentation: information that documents the relationships between the record creator, the record users and the record custodians. This can include information about associated records.

¹⁹ Andrew J. Flanagin and Miriam J. Metzger, "From Encyclopaedia Britannica to Wikipedia," *Information, Communication & Society* 14(2011): 258.

²⁰ Cook, "Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era," highlights the importance of context to provide meaning to archival records and to make visible the actions, agendas and functions which lead to and explain a record's creation. Cook also identifies the need for archivists both to recognize and to include the multiple contexts and understandings of archival records within archival description.

3. Control Documentation: information that documents the custodial history of the record. This can include control or other identifying numbers, transfer dates and agents, history of arrangement, and information pertaining to the creation, revision or deletion of retrieval tools.
4. Records Documentation: information that documents the chronology and use of the record. This should include the date of creation and specifics of reproduction and re-use.

User contributed content was considered contextual description for the purposes of this study, presuming that the content of the descriptive information fell into one of the documentation types outlined above. User contributed content consists of digital images, text, tags or any other form of content supplied by individuals external to the archival institution.

The research was conducted in two stages. Due to the exploratory nature of the research, the small number of subject websites, and the number of features on each to be examined, a case study methodology was used in Stage One. Attitudinal measurement was used in Stage Two. In Stage One, case study Web 1.0 and Web 2.0 sites were evaluated using a survey (scale) that measures features of the content design and execution in the representation of context. In Stage Two, user attitudes to Web 1.0 and Web 2.0 depictions of context online were evaluated using a Likert scale based on archivist and user cohorts.

3.1 Stage One: Research Design

In Stage One, to ensure consistency of evaluation, a Web 1.0 page was paired with the equivalent Web 2.0 page that documents the same record. Each pair of sites also represents a different example of Web 2.0 expression—a mash-up, a photo sharing site, and a wiki.

3.1.1. National Archives of Australia (NAA)

Web 1.0 – NAA Record Search: B2455, BUTCHER R T

The National Archives of Australia's "collection mainly documents Australian government activities since Federation in 1901."²¹ Its online catalogue, called RecordSearch, provides information about the records held by the Archives. Records are arranged hierarchically into functional series; item level, related records, and recording and controlling agency descriptions are provided via links at the series level description. Users to the site can perform basic or advanced keyword and subject searches or search by photograph, name or passenger arrivals index. The catalogue includes digital reproductions with some record descriptions, however the description was not accompanied by an electronic copy for the subject record of this paper.

Web 2.0 – Mapping our Anzacs: Ralph Thomas Butcher

Mapping Our Anzacs is a mash-up, participatory website managed and hosted by the National Archives of Australia. In 2007, the National Archives of Australia released online copies of all the records in Series B2455 which contained the records of the men and women who served in the armed forces during World War I. Mapping Our Anzacs provides access to these 375 971 service records as well as related user contributed content including biographies and photographs of veterans. Users to the site can add notes and photos to the virtual scrapbook or build a tribute to a loved one. Geographical searching is the primary

²¹ National Archives of Australia, "Using the Collection," <http://www.naa.gov.au/collection/using/>.

means of accessing the records. Users to the site can search by place of birth or enlistment or by veteran name.

The idea behind Mapping Our Anzacs is to use that place-based information to provide a new pathway to the records. We thought that a spatial pathway into World War I service records would make sense for local communities where, in many cases, a World War I memorial is central to the town and the community...It is the place-based information in each record title that give the records a structure. We have attempted to map each place name to a global position, so you can now find records by browsing the maps rather than by searching. You no longer need to know in advance what you are looking for, and you can see the relationships between records, rather than seeing each one in isolation.²²

Due to the arrangement of the records on the site and the method of enlisting and collecting user contributed content, the site is less focused on providing context to the service record; rather the service record and user content are made available primarily to provide context and biographical information about the individual military service person about whom the service record is written.

3.1.2 Library and Archives Canada (LAC)

Web 1.0 – LAC Archives Search: Dog Child, a North West Mounted Police scout, and his wife, The Only Handsome Woman, members of the Blackfoot Nation

Library and Archives Canada's online collections catalogue is called Archives Search. The catalogue provides descriptive and contextual information about the records held by Library and Archives Canada. Users to the site can search textual, photographic, iconographic, audio, philatelic, cartographic, architectural and other documents by performing a basic, advanced or image search. Basic searching is done by keyword, material type, hierarchical level and whether the item is digitally reproduced online. Advanced searching includes all basic search categories as well as date, and source fields. Image searching is by keyword and material type. Similar to The National Archives, records are arranged hierarchically into fonds or collections. Descriptive content is based upon the level being described and is not repeated at lower levels of description. Some descriptions include a link to a digital reproduction of the record.

Web 2.0 – Flickr: Dog Child, North West Mounted Police scout, and his wife, The Only Handsome Woman, members of the Blackfoot Nation

Library and Archives Canada also holds a pro account with the photo management and sharing site, Flickr. Flickr has over 10 million active groups and 60 million photographer users.²³ Library and Archives Canada uses the site to exhibit and make available digital reproductions of photographs from their archival collection. In many cases these photos are grouped into photostreams based upon a common category or theme. The subject item, a photograph entitled, *Dog Child, a North West Mounted Police scout, and his wife, The Only Handsome Woman, members of the Blackfoot Nation*, is grouped into the Pride and Dignity photostream. Images making up this photostream were originally part of an exhibition entitled, Aboriginal Portraits.

²² Mapping Our Anzacs, "About This Site," <http://mappingouranzacs.naa.gov.au/about.aspx>.

²³ Flickr, "Home," <http://www.flickr.com/>.

3.1.3 The National Archives (UK)

Web 1.0 – The Catalogue: Item CO 137/350/52

The National Archives' online catalogue contains "11 million descriptions of documents from central government, courts of law and other UK national bodies, including records on family history, medieval tax, criminal trials, UFO sightings, the history of many countries and many other subjects."²⁴ Its mandate is to make available information pertaining to the holdings of The National Archives. Information pertaining to the holdings is made available through keyword, date range and department or series code search capabilities. While the online catalogue provides important contextual information for the researcher about the Archives' holdings, the site does not contain digital reproductions of any archival documentation, thereby requiring the researcher to visit the Archives in person to view and use the records. Catalogue descriptions are arranged and described hierarchically with the department functioning as the highest level of description, followed by division, series, sub-series, sub sub-series, piece and item. The content of the description is based upon the contextual information at the level described. This information is not repeated at the lower levels of description. Catalogue descriptions are written by Archives staff.

Web 2.0 – Your Archives: Inside Kingston Lunatic Asylum: the case of Ann Pratt

One of The National Archives' web 2.0 initiatives was the wiki, Your Archives. The wiki was launched in 2007 as a means of "providing an online platform for users to contribute their knowledge of archival sources held by The National Archives and other archives throughout the UK."²⁵ Since that time over 31 000 people have registered and contributed or updated articles, over 21 000 articles have been created, almost 260 000 edits have been made, and there have been over 6 million visits to the site with more than 50 million page views.²⁶ User contributions to the wiki are intended to enhance the archival descriptions provided by Archives staff in The Catalogue, research guides, Documents Online and The National Register of Archives. Users can search or browse by the subject or content of the articles.

As redirection from a Web 2.0 site to a Web 1.0 finding aid is a common practice among archival institutions for providing fuller contextual descriptions, it was also necessary to identify whether the contextual information was presented through primary or secondary means. Primary presentation will be defined as information that is immediately available to the user on the Web 2.0 case study page. Secondary presentation will be defined as information that requires the user to perform an additional action, such as clicking on a link, in order to be viewed.

4. Stage One: Results

4.1 Presentation of Context

A review of the Web 1.0 sites and their Web 2.0 equivalents revealed a number of things about the presentation of context. In general, the mandate of each site determined the extent and nature of contextual information. Mapping Our Anzacs, for example, placed the soldier as the focus of the

²⁴ The National Archives, "The Catalogue," <http://www.nationalarchives.gov.uk/catalogue/>.

²⁵ Your Archives, "Home," http://yourarchives.nationalarchives.gov.uk/index.php?title=Home_page.

²⁶ Ibid.

description rather than the archival record presented on the site. As a result, the user contributed content is heavily biographical in nature. Regardless of the mandate of the site, there were some general commonalities in descriptive practice. None of the sites, either Web 1.0 or Web 2.0, provided significant amounts of contextual information through primary access. Within online catalogues this is likely due to the hierarchical descriptive practice within most archival traditions. As a result, much of the contextual information is not located at the record or item level, but instead can be found at the series or higher levels of description. In each Web 2.0 site archivists provide a hyperlink back to the item level of description in the online catalogue. By these means any user has access to fuller contextual descriptions should they be prepared to follow the required number of hyperlinks. The most common fields of description completed at the primary level, regardless of Web 1.0 and Web 2.0 environments, were control number (all sites save one—Mapping our Anzacs), date of creation (all sites), and creator name (excluding the two NAA sites). When this seemingly basic information is lacking, it may be due to descriptive tradition and the more biographical focus of the Web 2.0 site. The least common fields of description completed at the primary level were information pertaining to record control, specifically custodial information, such as transfer agents and dates. This may be due to archival descriptive tradition and the fact that the information is made available at a higher level. Specifics of reproduction, which one would expect to be described at the item level, were also excluded on most subject sites.

4.1.2 Creator Documentation

On most sites contextual documentation concerning records creators and creating agencies was limited to the name of the record's creator. All the sites, with the exclusion of the two National Archives of Australia sites, provided primary access to the name of the body responsible for the creation of the subject record. For the most part, primary access was not provided to additional information concerning the creator's function, mandate, personal history, or administrative or family structure. Only the Your Archives wiki provided personal background documentation concerning Ann Pratt, the author of the subject record.

4.1.3 Relationship Documentation

Only one online catalogue, that of the National Archives, provided primary access to information about associated records and this information was very limited in its scope, consisting of an arrangement structure with hyperlinks to the higher levels of description. The pamphlet by Ann Pratt is also mentioned at the item level in association with the government's response document, *Official Documents on the Case of Ann Pratt, the Reputed Authoress of a Certain Pamphlet...* The Library and Archives of Canada's online catalogue provided secondary access to documentation concerning associated records in a variety of ways, depicting a multiplicity of relevant contexts. The photographer's name is hyperlinked, providing the user with the ability of associating the subject photograph with others taken by the same creator. The 'show arrangement structure' link also allows users to associate the item with other records within the same fonds and series. A notation regarding the subject photograph's inclusion in an exhibition entitled, "Aboriginal Portraits," under the additional information field, not only provides users with the record's context of re-use but also documents new associations and relationships that have come about during LAC's custody of the record.

Documentation surrounding the context of association was more prominent on the Web 2.0 sites than in the online catalogues. All three Web 2.0 sites attempted to situate the subject record within a

wider scope of related records. This context of association, however, did not always mirror the associations provided by the online catalogue. The Mapping Our Anzacs site states that by mapping geographically and browsing maps rather than by searching, “you can see relationships between records, rather than seeing each one in isolation.”²⁷ While the Web 2.0 site links the records spatially, the NAA online catalogue links and associates records according to the office of creation, use and transfer. Similarly, LAC has situated its subject photograph within a folder entitled “Pride and Dignity.”²⁸ In both instances the Archives has, arguably, situated the record within an artificially constructed context that does not necessarily mirror the record’s original context of association. It may, however, reflect the record’s more modern context of use.

4.1.4 Control Documentation

Documentation concerning records control and custody both in online catalogues and in Web 2.0 was incomplete. All sites provided primary access to the record’s control number, but only The National Archives’ online catalogue provided primary access to additional control information. Evidence of the arrangement of the pamphlet within the larger record series is illustrated at the item level, though an explanation of the arrangement of this record and its associated series is only available at the higher levels of description through secondary access.

The majority of control documentation, such as evidence of arrangement, transfer agent and transfer date, is available at higher levels of description. In most cases this information is available to users of the online catalogues through secondary access via hyperlinking to the series or collection/fonds level descriptions. Access to this same control documentation is also available to users of the Web 2.0 sites via hyperlinks to the online catalogue, but the contextual information is that much more removed from the user, requiring multiple navigation actions on the part of the researcher.

4.1.5 Records Documentation

Records documentation, such as the date of creation, specifics of reproduction and record re-use, are not prevalent within either the online catalogues or the Web 2.0 sites. The date of creation was the only contextual information to be provided on all sites, with the exception of Mapping Our Anzacs.²⁹ There is some discrepancy regarding the date of the photograph, *Dog Child, a North West Mounted Police scout and his wife, The Only Handsome Woman, members of the Blackfoot Nation*, between the LAC online catalogue and the Flickr site. While the LAC catalogue lists the date of creation as ca. 1890, the Flickr site gives a more specific date of 1890. This may be the result of a conscious decision by LAC to simplify information for Web 2.0 users. An examination of other photographs within the same folder suggests that LAC is deliberately omitting the word “circa” from its Web 2.0 descriptions. Another photograph in the

²⁷ National Archives of Australia, “About This Site.”

²⁸ Pride and Dignity is an exhibition of over 60 photographic reproductions (c. 1846-c.1960) taken from the original exhibition Aboriginal Portraits from the National Archives of Canada produced by Edward Tompkins and Jeff Thomas, guest curator, and exhibited at Library and Archives Canada during the spring and summer of 1996.” See Flickr, “Pride and Dignity,” <http://www.flickr.com/photos/lac-bac/sets/72157624189241172/>.

²⁹ Although the date of record creation is present on the Mapping Our Anzacs site, the information was not supplied directly by the NAA. The original record was dated by the records creator. A digital reproduction of the item is available to view from the site thereby giving users access to the date of creation upon viewing the electronic copy.

same folder, *Innu (Montagnais) woman, probably taken at North West River, Labrador, ca. 1930*, has also had the “circa” removed from the date of creation field in Flickr, though the date in the online catalogue includes the word.³⁰ Regardless of the motivations behind the date change, the discrepancy provides two potentially different contexts of creation for the same record. If, in fact, the context was amended to simplify information for Web 2.0 users, it raises a number of concerns about the impact, and appropriateness of the simplification of context for Web 2.0 users.

It could be expected that since digital reproductions of the subject record are available on each of the Web 2.0 sites, some contextual information regarding reproduction and re-use would be more prominent on these sites. This does not appear to be the case, however, as none of these sites provide primary access to this information. Two of the three sites, Flickr and the Your Archives wiki, do provide varying degrees of contextual reproduction and re-use information at the secondary level. Your Archives contains two digital reproductions, one of the subject record and one of an associated record, each of which includes a file history field, available by clicking on the digital image. The file history records image specifics such as preview and full resolution size, file size and MIME type. It also indicates when the file was uploaded to the site, its dimensions and the creator or source of the reproduction. A note on the re-use of the LAC photo appears on the home page for the Pride and Dignity exhibit on Flickr.

Only LAC provides information about the re-use of its photo in its online catalogue. This documentation is available under the ‘additional information’ heading of the item description. Under this heading the photograph’s inclusion in an exhibition entitled, “Aboriginal Portraits” is mentioned.

4.1.6 User Contributed Content

Due to the diversity of sites chosen for this study, the nature of the user contributed content varied considerably between the sites. The wiki provided the contextually richest user contributed content, with users responsible for supplying the entirety of the record description. The mash-up and photo-sharing sites provided details specific to the subject and mandate of the site but did not necessarily add greatly to the context of the record.

The mash-up provided photos and biographical information which contributed to an understanding of the soldier about whom the record was written. While this biographical information is both interesting and important, it adds little to the context of the service record displayed on the page as the soldier was not responsible for the record’s creation and later use. It can be argued however that the biographical information has some contextual utility as it further authenticates the validity of the record.

Comments left by users on LAC’s Flickr page featuring the photograph of Dog Child also have limited contextual utility. The majority of the comments centre on the origins or type of sword that Dog Child is holding and how he came to acquire it. Similar to the mash-up, the additional information may be useful to further validate or invalidate the authenticity of the image, but it is not particularly contextually rich. It is notable that Flickr users also put forward theories as to how Dog Child acquired the sword. The users do not question the truth of the image before them, or that the image was, in any way, staged with props according to the photographer’s aesthetic. Instead they make guesses as to how Dog Child could

³⁰ Compare Library and Archives Canada, “Innu (Montagnais) Woman, Probably Taken at North West River, Labrador, Ca. 1930 / Femme Innu (Montagnais), Probablement Photographiée À North West River (Territoires Du Nord-Ouest), Vers 1930,” Flickr, <http://www.flickr.com/photos/lac-bac/4666289053/in/set-72157624189241172/>, with Library and Archives Canada, “Innu (Montagnais) Woman, Probably Taken at North West River,” http://collectionsCanada.gc.ca/pam_archives/index.php?fuseaction=genitem.displayItem&lang=eng&rec_nbr=3224646.

have acquired such an item. These guesses, which do not appear to be based upon any research, not only threaten to recreate a biographical history for Dog Child, but they also highlight users' unquestioning belief in the truth of the record presented to them.

4.2 Discussion

The results of the examination and comparison of the Web 1.0 and Web 2.0 sites suggests a number of trends regarding the presentation of context. First, archivists rely heavily on hyperlinks to furnish users with contextual information. Hyperlinking is used in both online finding aids and Web 2.0, though for arguably different purposes. Hyperlinking in Web 1.0 is used to provide access to higher levels of contextual description in order to facilitate a more complete contextual understanding and to eliminate the need for duplication of information between levels. By contrast, the practice of hyperlinking back to the online catalogue from Web 2.0 appears to be practised as a means of eliminating the need to provide anything more than a bare-bones description within Web 2.0. The expectation and reliance by archivists on users to seek additional contextual information through hyperlinking may not be justified, however. The Smithsonian's experience using Flickr suggests that there is little redirection by users from the Web 2.0 site to the archives' online catalogue.³¹

A second trend in the presentation of context is the difference in the contextual mandate between online finding aids in Web 1.0 and Web 2.0. In one study, Yakel observed that in different environments "archival information and records are uncontextualized, or at the very least differently contextualized."³² While her study noted the differences in descriptive practice between online and analogue finding aids, it is fair to suggest that a further difference exists between Web 2.0 description and that of online finding aids.

Online finding aids in Web 1.0 are more likely to focus on information associated with traditional ideas of context, such as the context of record creation, while Web 2.0 sites appear to be less concerned about the origins of the record and more focused on generating and presenting newer meanings. By grouping records geographically, in the case of the Mapping Our Anzacs site, or by subject, in the case of the Flickr site, the Archives are visually presenting a context, based not upon the origins of creation, but instead upon a consciously-curated mandate. While the context of creation remains accessible through hyperlinks, it is the newly-created context which is given the pride of place and predominance.

The relationship between context and trust was also highlighted through this site examination. It appears that users believe in the truth of what they see regardless of the amount of context provided by the archives. This suggests that while contextual information can be used to provide authenticity to a record, it is not necessarily a requirement for Web 2.0 users. Perhaps Web 2.0 users value archival records merely for their aesthetics and care little about their wider context. Schwartz notes that,

traditional item-level description of photographs, indexed by subject and credited to the photographer, but without adequate contextual information about their functional origins and provenance, or clear links to such contextual information, transforms photographic

³¹ Martin Kalfatovic et al., "Smithsonian Team Flickr: A Library, Archives, and Museums Collaboration in Web 2.0 Space," 272-273.

³² Elizabeth Yakel, "Impact of Internet-Based Discovery Tools on Use and Users of Archives" (paper presented at the XXXVI Roundtable on Archives (CITRA) Meeting, France, 2002), 195.

archives into stock photo libraries, reducing photographs to their visible elements and conflating photographic content and photographic meaning.³³

Alternatively, users may unquestioningly believe in the accuracy of the archival record simply because it was made available to them by a trusted archival repository. Yakel's research suggests that for researchers, "the presence and placement of the finding aid in the archives is an implicit sign of authority."³⁴

A final trend noted through this examination was the possible simplification of context on the Web 2.0 sites. This can be seen in the comparatively smaller descriptions provided on these sites as well as the removal of the word 'circa' from photograph dates on Flickr. While it is unclear why LAC chose to remove this word, its practice of doing so raises many questions. Is the online catalogue considered the only authoritative source of information and therefore all other arenas of access are not required to be accurate as long as they point back to the authority record in the online catalogue? If, in fact, the word 'circa' was removed as a simplification for Web 2.0 users, why is the word considered appropriate for the online catalogue? Without understanding the mandate of LAC's Flickr presence it is difficult to answer these questions. This practice does suggest, however, that not only do the Web 1.0 and 2.0 sites have different mandates, but it is expected that they will have different audiences, each of which requires a different type and quantity of contextual information. Duff and Harris state, "what we choose to stress and what we choose to ignore is always and unavoidably subjective, and the value judgements that archivists made affect in turn how our researchers find, perceive and use records."³⁵ The conscious decision to simplify, alter, or omit contextual information will affect users' understanding of archival records.

5. Stage Two: Research Design and Results

The second part of the research consisted of a quantitative analysis evaluating attitudes of archivists and patrons of archival services towards the contextual information presented on the case study sites. The target population consisted of professional archivists and users accessed from listservs and discussion lists in North America and Australia. A usable sample of 71 respondents comprising 40 Archivists ($n_1 = 40$) and 31 users ($n_2 = 31$) was constructed using snowball sampling. Data collection was carried out through the use of an online survey tool created through the Qualtrics online survey service. Each question had a link to the specific website to be evaluated. Questions were closed-ended and employed an ordinal scale (coded Strongly Disagree = 1; Strongly Agree = 5) to measure attitude. Question order was also randomized. Demographic questions were used to identify different user groups within the samples. The largest age cohort by percentage was 40-49 with a greater proportion of respondents in the younger age ranges (20-29, 30-39 and 40-49) as opposed to the older ranges. The most common level of education among respondents was a Master's Degree with a greater proportion of respondents possessing higher university degrees at the Master's and Doctorate levels. Variance within the samples was based upon familiarity with archival records and retrieval tools, and the type of archival research generally conducted. The two most predominant categories of research conducted by users were academic and genealogical.

³³ Joan Schwartz, "Coming to Terms with Photographs: Descriptive Standards, Linguistic 'Othering', and the Margins of Archivy," *Archivaria* 54(2002): 157.

³⁴ Elizabeth Yakel, "Archival Representation," *Archival Science* 3(2003): 13.

³⁵ Wendy Duff and Verne Harris, "Stories and Names: Archival Description as Narrating Records and Constructing Meanings," *Archival Science* 2(2002): 275.

The mode for use of archival facilities was 1-5 times in the past year with the predominant preferred method of accessing archival records being through a combination of both online and in-person visits.

Stage Two of the research tested the following hypotheses:

H₁: Attitudes to the *reliability* of contextual metadata are related to the host page type (Web 1.0 or Web 2.0).

H₂: Attitudes to the *reliability* of contextual metadata appearing on Web 1.0 and Web 2.0 pages are related to whether a person is an archivist or user.

H_{2.1}: Attitudes of archivists to the *reliability* of contextual metadata appearing on Web 2.0 sites differs from that of users.

H₃: Attitudes to the *use and reliability of comments and tags* to understand contextual metadata appearing on Web 2.0 pages are related to whether a person is an Archivist or user.

Survey participants were asked to respond to four statements regarding the reliability of the contextual data on the case study sites.

S1. The information on this page about the record's creator is reliable.

S2. The information on this page about when the record was created is reliable.

S3. The information on this page documenting how and why the record was created is reliable.

S4. I think the content supplied by users on this page is reliable.

For each statement, respondent data for each of the six case study sites was imported to SPSS with additional data describing the nature of the sites (Web 1.0 or Web 2.0) and the respondent type (archivists or users). The Wilks Shapiro statistic and a normality plot showed all data to be *abnormally* distributed.

5.1 Data Screening and Analysis: H₁

Overall respondents considered information about context reliable across Web 1.0 and Web 2.0 sites. The skew is much less pronounced for Web 2.0 across each of the statements, indicating that respondents are less positive about the reliability of contextual information describing the record's creator, when the record was created and how and why the record was created. This finding is most notable in the context of S3.

For each of the first three statements, the data was further explored using cross tabulation and a Mann-Whitney U Independent Samples test. Across S1, S2 and S3, the cross tabulation showed that both archivists and users regard Web 1.0 as more reliable than Web 2.0. Statistically significant differences at the 95% ($\alpha=0.05$) confidence level were demonstrated with statements S1 ($p=.000$), S2 ($p=.000$), and S3 ($p=.031$). These differences were significant enough to reject the H₁ null hypothesis in relation to statements S1, S2 and S3.

5.2 Data Screening and Analysis: H₂

Further testing was conducted to see if significant difference existed in attitude between archivists and users. Generally, respondents considered the information presented to be reliable. The skew is significantly less pronounced among both archivists and users for S3, suggesting that both groups were less positive about the reliability of information concerning how and why the record was created. Additionally, a markedly higher skew amongst archivists' in response to S2 indicates that they were more positive than users about the reliability of information about when the record was created.

For each of the first three statements, the data was explored using cross tabulation and a Mann-Whitney U Independent Samples test. Statistically significant differences at the 95% ($\alpha=0.05$) confidence level were demonstrated with statement S2 ($p=.028$) resulting in a rejection of the H_2 null hypothesis for this statement. The low p -value of S1 ($p=.069$), while not statistically significant, suggests that further study, employing a larger sample size, may be indicative of an attitudinal difference between groups. Overall, users are less positive towards the reliability of contextual information, particularly information pertaining to who and when the record was created (S1 and S2 respectively) than are archivists. There is no marked difference in attitude between archivists and users towards the reliability of contextual information pertaining to how and why the record was created (S3).

Additional Mann-Whitney tests were conducted on all three statements to test $H_{2.1}$. Statistically significant differences at the 95% ($\alpha=0.05$) confidence level were not demonstrated with any of the statements. Therefore $H_{2.1}$ is rejected for all three statements. The p -value for S1 ($p=.055$) and S2 ($p=.061$) do suggest the possibility that there may be a difference in attitude between archivist and users to the reliability of contextual information pertaining to the record's creator and when the record was created on Web 2.0 sites. A larger sample size is necessary to evaluate this possibility further.

5.3 Data Screening and Analysis: H_3

Additional testing was carried out to determine if significant difference existed in attitude between archivists and users to the reliability of user contributed content. The data displays a negative skew, indicating agreement among both archivists and users that comments and tags are reliable. The more pronounced negative skew and higher mode among users suggests that they are more positive about the reliability of comments and tags than are archivists.

The cross tabulation shows that the majority of both archivists and users neither agree nor disagree that comments and tags are reliable. Few respondents display strong agreement with the statement and both groups have some variation in attitude among respondents regarding the reliability of comments and tags. Statistically significant differences at the 95% ($\alpha=0.05$) confidence level were not demonstrated with statement S4 ($p=.728$) resulting in a retention of the H_3 null hypothesis. The low p -value of S1 ($p=.069$), while not statistically significant, suggests that further study, employing a larger sample size, may be indicative of an attitudinal difference between groups. Overall, users are less positive towards the reliability of contextual information, particularly information pertaining to who and when the record was created (S1 and S2 respectively) than are archivists. There is no marked difference in attitude between archivists and users towards the reliability of contextual information pertaining to how and why the record was created.

5.4 Discussion

Statistical testing reveals that both archivists and users are inclined to find the information on a Web 1.0 site to be more reliable than information found on a comparable Web 2.0 site. In addition, there is evidence to suggest a difference in attitude to the reliability of contextual information between archivists and users. Specifically, archivists are more likely to find contextual information about the record's creator and when the record was created to be reliable than users. While no statistically significant differences were noted between the attitudes of archivists and users to the reliability of contextual information on Web 2.0 sites, the results of statistical testing do suggest the possibility that a difference in attitude may

exist. Further testing, employing a larger sample size, is required to determine the extent of attitudinal differences between archivists and users and whether this difference is equally present amongst Web 1.0 and Web 2.0 sites.

Statistical testing also reveals that both archivist and user groups are not strongly disposed to agree or disagree that comments and tags are reliable. There was some variation of response among both groups of respondents indicating that there is little consensus of opinion regarding the use of comments and tags as contextual information.

Overall, the perceived reliability of contextual information is significantly different between Web 1.0 and Web 2.0. This may be due, in part, to users' reluctance to trust user contributed content and sites which contain this type of information. Flanagin and Metzger have noted that perceived low-expertise sources are considered less credible and therefore less trustworthy.

The lack of author identification makes it difficult to determine whether information is biased, since users cannot know the motives for information provision, and the lack of cues about the expertise of contributors similarly inhibits users' capacity to determine the accuracy of information provided.³⁶

The inherent lack of credibility of user contributed content, irrespective of the content on the specific survey case study sites, was also suggested by the survey results, which indicated that the Your Archives wiki was not considered reliable, despite the fact that its content was written predominantly by a National Archives archivist. Survey respondents, both archivists and users, were neutral towards the reliability of user contributed content as contextual information and this suggests the possibility that archivists do not consider contextual information on Web 2.0 to be as authoritative as information in the Web 1.0 finding aid. These results, particularly among archivists, are surprising given the scholarly recognition of the potential of user contributed content to broaden and enhance record descriptions.

6. Conclusion

A review of relevant literature has indicated that authors have written extensively on the fact that context is integral to the evaluation of archival records for authenticity and the establishment of user trust. Context has been recognized as incorporating not only the original provenance of creation but also subsequent interactions and use by both users and archivists.³⁷ In a digital environment, context becomes particularly important as researchers must evaluate a copy of the original record, often in isolation from contextual clues such as physical associations between records and documentation of digital reproduction.

³⁶ Flanagin and Metzger, "From Encyclopaedia Britannica to Wikipedia," 258.

³⁷ See James Opp, "The Colonial Legacies of the Digital Archive: The Arnold Lupson Photographic Collection," *Archivaria* 65(2008): 3-19; Koltun, "The Promise and Threat of Digital Options in an Archival Age"; Millar, "The Death of the Fonds and the Resurrection of Provenance: Archival Context in Space and Time"; Frank Upward and Sue McKemmish, "Somewhere Beyond Custody: Literature Review," *Archives and Manuscripts* 22(1994): 136-149; Yakel, "Archival Representation"; Duff and Harris, "Stories and Names: Archival Description as Narrating Records and Constructing Meanings"; Elisabeth Kaplan, "'Many Paths to Partial Truths': Archives, Anthropology, and the Power of Representation," *Archival Science* 2(2002): 209-220.; Huvila, "Participatory Archive: Towards Decentralised Curation, Radical User Orientation, and Broader Contextualisation of Records Management"; and Terry Cook, "Archival Science and Postmodernism: New Formulations for Old Concepts," *Archival Science* 1(2001): 3-24.

The mandate and structure of Web 2.0 adds a further level of complexity to the nature and evaluation of contextual information. While the opportunity exists for user contributed content to enhance context by permitting multiple voices and interpretations of the record to be recognized, Web 2.0 is largely unmediated, allowing both accurate and inaccurate information to exist side by side, thus blurring, rather than enhancing, contextual information.

The intent of this study was to examine the differences in the presentation of context between Web 1.0 and Web 2.0 sites and measure differences in attitude towards the reliability of this information among archivists and users. Though only a preliminary study, the study found variation in the way in which institutions described context online and significant differences in attitude towards the presentation of context on Web 1.0 and Web 2.0.

It is clear, both from the site analysis and attitudinal survey, that the presentation of context, and attitudes towards the reliability of this information, are significantly different between Web 1.0 and Web 2.0. Despite the opportunities that Web 2.0 provides for broader, richer, contextual information, archivists do not appear to place much stock in the reliability of this information. Perhaps it is this distrust of user contributed content among archivists that has resulted in Web 2.0 sites serving a different mandate than their Web 1.0 counterparts, with Web 2.0 sites functioning less as sources of contextual information and more as outreach and awareness tools—places where select archival collections can be highlighted and new contexts and associations created. Through the creation of artificial contexts or displacement of original contexts, archivists are changing the contextual framework within which records will be evaluated. The resulting lack of contextual clues and the online search tools and item-level tagging and description threaten to present archival records in isolation without both physical and intellectual clues as to their history and reason for inclusion online, weakening the perceived reliability of the contextual information presented and ultimately undermining the authenticity and integrity of the archival records.

Preserving Social Media

Opening a Multi-Disciplinary Dialogue

Lisa P. Nathan¹ and Elizabeth Shaffer²

School of Library, Archival and Information Studies, iSchool @ The University of British Columbia, Canada

¹*Lisa.Nathan@ubc.ca*; ²*eshaffer@mail.ubc.ca*

Abstract

Digital artefacts generated through use of social media tools have potential long-term value to individuals, organizations and societies. If there is a desire to systematically collect and preserve accounts of daily life, government activities, and societies' documentary heritage, archival approaches must account for changing information systems—the tools, policies, and practices through which we engage in the contemporary information ecosystem. Through this paper we argue that in light of the growing complexity of digital information practices, particularly in relation to the use of social media, archivists need look to the scholarship of design and planning, in particular the work of human computer interaction designers. In turn, the designers of digital information systems need to engage, draw upon, support, challenge and inform contemporary archival theory and practice.

Authors

Lisa P. Nathan is an Assistant Professor at The University of British Columbia's School of Library, Archival and Information Studies (the iSchool@UBC). Since 2010, she has served as the Coordinator for the school's First Nations Curriculum Concentration. Her research is motivated by the high potential for interactions with information systems to have a long-term influence on the human condition. Through a range of projects she investigates theory and method related to the design of information systems that address long-term societal challenges; information practices that develop and adapt as we use these systems; and factors that influence the sustainability of these systems over time.

Elizabeth Shaffer is a doctoral candidate at the School of Library, Archival and Information Studies (the iSchool@UBC) from which she received a Master of Archival Studies degree. Her dissertation research is at the intersection of social media and archival theory focusing on the preservation, policy and recordkeeping challenges posed by information and communication technologies, in particular social media. Her work aims to build on existing archival theory and practice to inform policy on social media use and records management and preservation.

1. Introduction

The digital artefacts generated through use of social media tools have potential long-term value to individuals, organizations and societies. If there is a desire to systematically collect and preserve accounts of daily life, government activities, and societies' documentary heritage, archival approaches must account for changing information systems—the tools, policies, and practices through which we engage in the contemporary information ecosystem. Through this paper we argue that in light of the complexity of digital information practices, particularly in relation to the use of social media, archivists need look to the scholarship of design. In turn designers, specifically designers of digital information systems, need to engage, draw upon, support and challenge contemporary archival theory and practice. Through this paper we argue that the field of human computer interaction is uniquely positioned to work with archivists to both inform archival theory and practice and in turn to be informed by archival theory and practice.

Archivists are well aware of the long-term influence records and the information systems that hold them have on the ability of future generations to access and interrogate their documentary heritage. We illustrate the potential for these collaborations by applying the design scholarship of Rittel and Webber¹ to a specific archival challenge, governments' use of social media. Through a discussion of Rittel and Webber's concept of a "wicked problem" we identify paradoxes that arise as system designers, users, and archivists approach social media artefacts from a plurality of conceptions concerning what is of value. We identify intersections where these interested parties might engage the problematic situations that the wish to preserve social media documentary objects foregrounds. We posit that this nascent conversation between concerned parties has the potential to lead to future generations of digital platforms that more deeply engage the broad societal challenges of the long-term preservation of digital documentary artefacts.

2. Context

The near ubiquitous use of social media by individuals, organizations and governments is generating documentary digital artefacts, the nature of which much is unknown. The various platforms individuals and organizations utilize to generate social media artefacts are often held by third party for-profit organizations whose business models are predicated on the collection, aggregation and monetization of users' data. The ephemeral nature of this web based information and the rapid evolutionary nature of social media tools and technologies facilitates an information ecosystem whose modalities, affordances and practices challenge the application of traditional archival functions such as appraisal, preservation and access. Just in the past five years we witnessed the disappearance of entire online communities and the majority of their digital artefacts with the dissolution of GeoCities and the rebranding of the social bookmarking site delicious when it was purchased by Yahoo in 2005. The business models that facilitate the use of and access to these sites and their data are often in direct conflict with the obligations and motivations of societies to ensure the preservation of their digital heritage. Individuals, organizations and governments utilizing these platforms are challenged because of the conflicting values present in this environment—those of the platforms' designers; the platforms' owners; the platforms' users and the social, political and technical environments in which these roles operate.

These examples highlight how the challenge of preserving digital heritage increases exponentially as ever shifting digital information systems evolve, along with the practices and policies through which we enact them. Yet, the *UNESCO Charter on the Preservation of Digital Heritage*² declares the necessity to ensure the long-term preservation of the world's digital heritage in all forms. It states that digital heritage is a "common heritage" amongst all nations consisting of assets that are "unique resources of human knowledge and expression"—many of which "have lasting value and significance" requiring protection and preservation for future generations. The purpose of which is the long-term accessibility to the public. The threat of loss is very real and action is needed by all stakeholders as "continuity of the digital heritage is fundamental."³ The development of social media is simply one of many phases in the continuing evolution of information and communication technologies, and based on current tools and

¹ Horst W. J. Rittel and Melvin M. Webber, "Dilemmas in a General Theory of Planning," *Policy Sciences* 4, no. 2 (1972): 155-169.

² UNESCO, 2003, http://portal.unesco.org/en/ev.php-url_id=17721&url_do=do_topic&url_section=201.html.

³ Ibid.

practice, we are likely to lose the documentary artefacts from this stage. Although loss of entire systems of knowledge is by no means unique in the history of humankind, we are not convinced that this is a precedent to intentionally follow.

3. Reframing by Design

Through this position paper we argue that if the long-term preservation of documentary digital artefacts created through engagements with social media, is a goal, it requires the archival profession to (re)consider *where* in the digital artefact generation process they engage. Similar to the shift that happened in the realm of electronic records, when it became necessary for archivists to inform the design and use of formal record systems, we propose that archivists engage proactively with those who design contemporary and future information systems. As a nascent step in this direction, we suggest an inquiry into the field of design where scholars are well versed in the art of taking action in problematic situations when there are many unknowns.

To start the inquiry, we turn to the work of Rittel and Webber in the early 1970's and their description of a "wicked problem."⁴ Working at the intersection of the design and planning disciplines, Rittel and Webber's work is in large part a response to growing public dissatisfaction with unsuccessful crime reduction programs, disappointing education reform projects and similar failed societal initiatives. They begin with a strong critique of the idealized, linear representations of societal problems prevalent in the early 70's. They illustrate why rationalistic, hard science inspired approaches to highly problematic situations are doomed in large part because societal problems are by their very nature multi-faceted and ever changing, thus they are neither fully knowable nor solvable. As an alternative, Rittel and Webber introduce the term wicked problem, a moniker for societal problems that "are never solved. At best they are only re-solved-over and over again."⁵

We posit that if there is a desire to systematically collect and preserve accounts of daily life, government activities, and societies' documentary heritage as generated within social media platforms, then striving to meet that desire represents a *wicked problem*. In the following paragraphs we apply select characteristics of wicked problems as described by Rittel and Webber to one type of social media engagement, governments' communications with their publics. Through this exercise, the government adoption of social media serves as an exemplar to illustrate the potential of design to help archivists reframe and engage challenge of "preserving" social media documentary artefacts.

In the following section we discuss select features of wicked problems and demonstrate how these characteristics can be fruitfully applied to the context of governments adopting social media tools. There are more features to wicked problems, but we wish to be illustrative, not exhaustive. We explicate Rittel and Webber's points, drawing them into a government context in an effort to ground the higher level argument, that the design of information communication technologies (e.g., social media tools) needs to be informed by the scholarship and discourse around the long-term access to our digital heritage.

⁴ Rittel and Webber, "Dilemmas," p. 160.

⁵ Ibid.

3.1. No Definitive Formulation

Rittel and Webber argue that “the formulation of a wicked problem *is* the problem!”⁶ It is necessary to formulate the problem and conceive of the solution at the same time. Their presentation of the hard sciences approach describes a linear process from problem statement, to hypothesis, to experimentation and then to solution. However, for wicked problems, they claim that through determining the right question, one begins to develop the solution. Consider the question: how to preserve trustworthy digital artefacts generated with social media over space and time. A positivistic approach would start with identifying the problem, then moving through a series of steps to figure out a solution. For wicked problems, the model of planning is that of an “argumentative process in the course of which an image of the problem and of the solution emerges gradually among the participants, as a product of incessant judgment, subjected to critical argument.”⁷

Governments have harnessed the affordances of social media technologies to open up innovative communication channels to facilitate immediate and ongoing interaction with their citizens.⁸ This engagement often utilizes a variety of third-party platforms with any number of evolving affordances. Yet, little is fully understood about the attributes of these products, the potential challenges they pose to archival theory and practice and whether it is possible to effectively apply current concepts of archival theory and practice to ensure that social media artefacts contribute to a trustworthy documentary heritage.⁹ Additionally, the policy environment within which these platforms are utilized often predates the onset of these technologies and has its foundation in a pre-social media environment. It is early in the investigation into the complexity of social media, the artefacts generated, the online environment they inhabit and the challenges they pose to traditional archival theory and practice. Archival scholars, practitioners and organizations are in the nascent stages of the “argumentative process” and an “image of the problem” is manifesting from the dialogue around the potential issues that social media artefacts and their long-term preservation present.¹⁰ Additionally, ideas and approaches are appearing in the literature, representing initial dialogues seeking (re)solutions to issues of appraisal,¹¹ capture and preservation,¹² and

⁶ Ibid., p. 161.

⁷ Ibid., p. 162.

⁸ Paul T. Jaeger, Jimmy Lin, J., and Justin M. Grimes, “Cloud Computing and Information Policy: Computing in a Policy Cloud?” *Journal of Information Technology & Politics* 5, no. 3 (2008): 269-283.

⁹ Michael Moss, “Without the Data, the Tools are Useless: Without the Software, the Data is Unmanageable,” *Journal of the Society of Archivists* 31, no. 1 (2010): 1-14.

¹⁰ Luciana Duranti and Elizabeth Shaffer, “eLearning Records: Are There Any to Manage? If So, How?” in *Social Media and the New Academic Environment: Pedagogical Challenges*, ed. Bogdan Patru, Monica Patrut, and Camelia Cmeciu (Vancouver: IGI Global, 2013); Patricia Franks, “Understanding Web 2.0 and Challenges for the Records Manager,” *Information and Records Management Annual* (2009): 107-121, <http://www.businessofgovernment.org/report/how-federal-agencies-can-effectively-manage-records-created-using-new-social-media-tools>; Bruce W. Dearstyne, “Blogs, mashups and wikis oh my!” *The Information Management Journal* 41, no. 4 (2007): 25-28, 30, 32-33; Sharon Henhoeffter, *Web 2.0 and Recordkeeping: Context and Principles* (Ottawa: Library and Archives Canada, 2010), <http://www.collectionscanada.gc.ca/digital-initiatives/012018-3401-e.html>; Theresa Smith, “#History #Dilemma: Social Media Proving Difficult for Archivists,” *Postmedia News* (July 13, 2012); Leisa Gibbons, “Testing the Continuum: User-Generated Cultural Heritage on YouTube,” *Archives and Manuscripts* 37, no. 2 (2010): 89-112.

¹¹ National Archives and Records Administration (NARA), and National Records Management Program (NRMP), *A Report on Federal Web 2.0 Use and Record Value*. 2010, <http://www.archives.gov/records-mgmt/resources/web2.0-use.pdf>; Christopher A. Lee, “Collecting the Externalized Me: Appraisal of Materials in the Social Web,” in *I*,

policy.¹³ Those engaging in the discourse around these various projects are attempting to articulate the problem space in order to identify potential (re)solutions.

3.2. No Stopping Rule

Rittel and Webber posit that wicked problems can never be solved for good and one can always do better, however, factors that are external to the problem necessitate stopping when a “good enough” (re)solution is achieved.¹⁴ Such factors may be limited resources—time, money, or political will. There is no singular “solution”; rather, there is a continuous series of ever-evolving processes based on a combination of theory and empirical inquiry that inform the design of systems, practices and procedures.

Governments are rapidly adopting social media¹⁵ predicated on the belief that it is an effective way to engage with citizens through dissemination of information and active, multi-directional communication.¹⁶ As such, the adoption often follows a bandwagon approach, utilizing the most prevalent and popular tools at hand. This approach points to the no stopping rule as archivists are required to spend time, resources and resolve to continuously develop their knowledge of new platforms, and accompanying affordance and products. Additionally, external factors such as limited resources and regulatory and legal constraints in a government context contribute to the constant need to find “good enough” solutions.

3.3. Solutions are neither true nor false

Wicked problems have no true or false answers. The assessment of proposed solutions are often judged on criteria of “better or worse” or how “satisfying” a solution may be.¹⁷ Government use of social media creates a variety of influences that will affect the satisfaction of any given solution. Tensions between and among legislation governing privacy, freedom of information, data protection, intellectual property, access and retention all complicate a potential agreed upon set of formal rules to determine the correctness of a solution. (Re)solutions that are supported technically may not fully support ethical or legal responsibilities.

Digital: Personal Collections in the Digital Era, ed. Christopher A. Lee (Chicago: Society of American Archivists, 2011), 202-240.

¹² National Archives and Records Administration, “Implications of Recent Web Technologies for NARA Web Guidance,” <http://www.archives.gov/records-mgmt/initiatives/web-tech.html>; Brand Neimann, “A Gov 2.0 Spin on Archiving 2.0 Data,” *Federal Computer Week* (January 27, 2011), <http://fcw.com/articles/2011/01/31/comment-brand-niemann-social-media-archives.asp>; Jennifer Wright, “To preserve or Not to Preserve: Social Media,” *The Bigger Picture* (June 13, 2012), <http://siarchives.si.edu/blog/preserve-or-not-preserve-social-media>; Gibbons, “Testing.”

¹³ Franks, “How Federal Agencies Can Effectively Manage Records Created Using New Social Media Tools,” *IBM Center for the Business of Government*, 2010, <http://www.businessofgovernment.org/report/how-federal-agencies-can-effectively-manage-records-created-using-new-social-media-tools>; American Council for Technology (ACT), and Industry Advisory Council (IAC), “Best practices Study of Social Media Records Policies,” 2011, <http://www.egov.vic.gov.au/website-practice/web-2-0-a/social-networks-and-social-media-in-government/best-practices-study-of-social-media-records-policies-in-pdf-format-646kb.html>; Jana Hrdinova, Natalie Helbig, and Catharine Peters, *Designing Social Media Policy for Government: Eight Essential Elements* (Center for Technology in Government: The Research Foundation of State University of York, 2010).

¹⁴ Rittel and Webber, “Dilemmas,” p. 162.

¹⁵ Mohammed Dadashzadeh, “Social Media in Government: From eGovernment to eGovernance,” *Journal of Business & Economics Research* 8, no. 11 (2010): 81-86; Henhoffer, *Web 2.0*; Jaeger et al., “Cloud Computing.”

¹⁶ Omar El Akkad, “Clement Pushes Twitter Integration in Redesign of Government Websites,” *The Globe and Mail*, November 29, 2011; F. Diane Lux Wigand, “Adoption of Web 2.0 by Canadian and US Governments,” in *Comparative e-Government*, ed. Christopher G. Reddick (New York: Springer, 2010), 161-181.

¹⁷ Rittel and Webber, “Dilemmas,” p. 163.

3.4. There is no immediate and no ultimate test of a solution to a wicked problem

Experimentation with dependent and independent variables that allow a scientist to determine causal relationships are not possible when one is wrestling with a wicked problem. The variables refuse to be isolated. It is not possible to trace all of the repercussions of complex information systems (tools, practices, policies, protocols) developed to preserve social media, to a singular, temporal stopping point where they can be judged. There are unforeseen consequences to each solution that are played out over an “unbounded” period of time.¹⁸ These dynamic consequences, in a government environment, can lead to undesirable outcomes that have impacts across contexts, including policy, decision making, and/or legislation.

3.5. No true trial and error period

With wicked problems, every implemented solution is consequential and the traces of these solutions cannot be undone. A system’s very existence is going to influence human lives resulting in various personal, organizational, and political repercussions. A government’s use and subsequent action to engage with citizens via social media does not create “trial” information and/or records, but rather potential digital heritage material with long-term value. If the choice of system allows the monetization of citizen data, the design of data within a system wherein the identity or integrity of the information is in question, or if users’ privacy and security are put at risk by using the system, such actions have irreversible repercussions. Subsequently, the actions taken by governments to preserve this information, if implemented with an ineffective solution, could result in the loss of irreplaceable documentary heritage and/or potential records of government actions and accountability. As such, every attempted solution counts and must be weighed both for its repercussions and the consequences that may result from the necessity to “undo” the solution. For example, a government’s decision to elicit input into policy making or elicit voter registration via a third-party social networking site such as Facebook¹⁹ are actions with repercussions for the potential preservation of and access to this information. Can the government ensure the trustworthiness of data collected via Facebook? If this is the only way to engage with the process, what are the implications of citizens being “required” to use Facebook? What are the public policy implications for data access, privacy, and freedom of information?

Finally, we reflect on Rittel and Webber’s articulation of why they chose the term wicked:

[W]e are calling them ‘wicked’ not because these properties are themselves ethically deplorable. We use the term ‘wicked’ in a meaning akin to that of ‘malignant’ (in contrast to ‘benign’) or ‘vicious’ (like a circle) or ‘tricky’ (like a leprechaun) or ‘aggressive’ (like a lion, in contrast to the docility of a lamb).²⁰

Although a wicked problem is not indicative of malicious intent, it may be morally questionable, even deplorable, to ignore it. The design and use of social media by government bodies is by no means a morally deplorable act. It is actually necessary for citizens to gain access to their government through initiatives such as Open Data and Open Government. However, to ignore the difficulties of preserving the

¹⁸ Ibid.

¹⁹ Cyrus Farivar, “Washington State Will Enable Voter Registration via Facebook,” *Ars Technica* (July 17, 2012), <http://arstechnica.com/business/2012/07/washington-residents-to-be-able-to-register-to-vote-via-facebook/>.

²⁰ Rittel and Webber, “Dilemmas,” p. 160.

documentary artefacts created through these interactions with citizens has strong ethical (and legal) implications.

Through the exercise of applying the design concept of a wicked problem to the issue of preserving the documentary artefacts created through governments' use of social media to communicate with its citizenry, we develop an appreciation for the insights of design scholarship. The inquiry highlights the lack of a definitive formulation of the problem and that the argumentative process, the ongoing debates about what the problem is and how to address it, is indeed a critical part of the process. In turn, it is a process with no stopping point in that the challenges of digital preservation will never be solved for good. Instead there will likely be a series of iterative "good enough" (re)solutions. Indeed, the wicked problem framing suggests that there is no ultimate test for these solutions, no trial and error period, and there are many contextual constraints.

All of these obstacles, and yet, action needs to be taken. We suggest that archivists expand the dialogue to include consideration of design perspectives, not just in terms of utilizing design theory, but through direct engagement with the designers of digital information technologies. As one potential area for collaboration, consider the field of human computer interaction. Human computer interaction researchers and practitioners are informing, if not directly designing, the digital platforms upon which social media interactions take place and the resulting artefacts they generate.

4. Archivists Informing System Design

Scholars and practitioners from the field of HCI can assist archival theorists in facilitating the interrogation of attributes of records within a social media context, and in turn, be informed by archival theory in the design of future information systems. With a deep knowledge of human cognition, technological capabilities, networking, human computer engagement and the importance of cultural contexts, the field of HCI is particularly well positioned to buttress archival goals in the social media milieu. However, they must first develop an understanding of archival goals and the centrality of the record.

Archives have the responsibility of safeguarding and preserving *records* of citizens, institutions and governments for use by current and future generations. Within archival theory and practice there exists a nuanced understanding of the attributes of a record. Although contemporary social media documentary artefacts appear to hold the promise of serving as records they may not hold all the attributes of a traditional record. For them to do so presents complex challenges due in large part to their form and the context in which these forms of information are created. However, what these forms of information do hold is the potential for informing societal memory and holding governments, institutions (public and private), and individuals accountable for their actions.

As Terry Eastwood points out, "the first object of archival theory is the nature of archival documents or records."²¹ Archivists' expertise lies with the creation, preservation and use of records and archives and the context of their creation and use—including the social, legal, technological and cultural environment.²² The ability of the document to attest to the fact and act it captures for action, future reference and/or to extend memory,²³ the trustworthiness of records is what archivists can contribute to

²¹ Terry Eastwood, "What is Archival Theory and Why is it Important?" *Archivaria* 37 (1994): 122–130.

²² Elizabeth Shepherd, "Archival Science," in *Encyclopedia of Library and Information Sciences*, ed. Marsha J. Bates and Mary Niles Maack, 179–191 (Taylor & Francis, 2010).

²³ Eastwood, "Archival Theory."

the broader discourse on systems design. How might archival concepts such as trustworthiness and archival practices inform/educate designers as they work to design systems?

Politicians hold town halls and announce policy reform,²⁴ governments encourage voter registration²⁵ and citizens engage with agencies and commissions²⁶— all within third-party social media platforms. All of these actions have the potential to generate digital artefacts that contribute to the documentary heritage of the early 21st century. Yet, the modalities and ephemeral nature of social media artefacts can be antithetical to the archival requirement to preserve authenticity and ensure reliability and accuracy over time and space. It is not enough to “save it all”—archival attention to provenance and context, archivists’ understanding of the concepts that underpin trustworthy records are relevant principles that should be applied in this shifting social media landscape in order to effectively capture and preserve this digital documentary heritage.

At this point in the conversation we are not cheeky enough to suggest that we have answers; however, we have plenty of questions such as: how can archival theory and practice evolve to keep pace with rapidly changing technologies and transforming information practices in contexts utilizing social media? How can archival theory and practice inform the design of social media information systems? How can the information policies that mediate and regulate information practice account for ever-evolving technologies and their affordances?

5. Conclusion

Current practices in the field of HCI focus on ways to support interaction in the shorter-term rather than issues of longer-term preservation, societal heritage and understanding. In the past, thinking about how to preserve records post hoc worked pretty well. The difference between preserving records post-hoc in the pre-digital environment and the complexity of trying to do the same in the current information and communication technology environment is enormous. Pre-digital it was possible to gather documents in a box at semi regular intervals, knowing that at a later date it would be possible to effectively preserve the documents deemed records when resources permitted. In a paper environment, there was often sufficient information recorded on the documents themselves and within the recordkeeping system, information practices were often understood well enough to reconstruct context, and the medium was stable as long as environmental controls were monitored. In the contemporary social media environment, there is no digital box that Facebook transactions are being tossed into. We cannot depend upon standardized information being encoded in these transactions, the information practices are not well understood, there is limited or no control over the creation of these documents, and the medium is less stable than invisible ink. As research into long-term digital preservation has shown, preservation in the digital environment must begin at creation.²⁷ In the born digital environment archival input needs to be part of the initial information system design. If an archival professional is called upon only after the systems are created and the practices and policies are in place, all she can do is provide a professional opinion of what has been and will continue to be lost.

²⁴ Smith, “Social Media.”

²⁵ Farivar, “Voter Registration.”

²⁶ Truth and Reconciliation Commission of Canada, <http://www.trc.ca/websites/trcinstitution/index.php?p=3>.

²⁷ International Research on Permanent Authentic Records in Electronic Systems (InterPARES), www.interpares.org.

The Role of Culture in Digitization and Digital Preservation

Preservation Cultures

Developing a Framework for a Culturally Sensitive Digital Preservation Agenda

Fiorella Foscarini, Gillian Oliver, Juan Ilerbaig and Kevin Krumrei

Abstract

All current attempts to frame and normalize organizational practices in the digital domain have certainly enriched our understanding of contemporary recordkeeping. However, they have also functioned as mechanisms that have helped reduce rather than explore complexities. When tested against real-world situations, existing models and standards only appear to work in relation to types of organizations that resemble traditional bureaucracies. Additionally, because they are designed according to some engineering-like approach, those models prove to be hardly applicable to ill-defined problems, as most of the problems faced by information specialists tend to be. This paper argues for a non-prescriptive, empirical, and situated approach to digital preservation, one that takes into account the cultural properties (i.e., values and practices) of any stakeholder involved. Investigating cultures at all levels (i.e., organizational, national, supra-national, professional, group) and, in particular, trying to understand 'information cultures' (i.e., the values accorded to information and attitudes towards it) as well as the factors affecting those cultures, are the main goals of a research project whose initial steps are described in this paper. Applying the information culture concept will enable the development of a culturally sensitive framework for digital preservation.

Authors

Fiorella Foscarini is an Assistant Professor in the Faculty of Information at the University of Toronto, Canada. She holds a Ph.D. in archival studies from the University of British Columbia in Vancouver. Before joining academia, she worked for a decade as senior archivist for the European Central Bank in Frankfurt am Main, Germany. In her teaching and research, Fiorella uses diplomatics, genre theory, and organizational culture concepts to explore the meaning of records creation, and to enrich the theory and practice of records management.

Gillian Oliver is a Senior Lecturer in Archives and Records Management at Victoria University of Wellington, New Zealand. Her doctoral degree is from Monash University in Australia. Her professional practice background includes developing the initial requirements for Archives New Zealand's digital archives. Gillian's research interests centre on understanding organizational dynamics, and the influences of culture on behaviours and attitudes to information.

Juan Ilerbaig has a Master's in Information Studies from the iSchool, University of Toronto (2012), and a Ph.D. in the History of Science and Technology from the University of Minnesota (2002). His research interests focus on the intersection between record keeping and the practice of science, with particular attention to such issues as communication genres and information culture. He has published articles in the *American Archivist* and the *Journal for the History of Biology*, among other journals.

Kevin Krumrei is a graduate student in the iSchool at the University of Toronto. Having completed graduate work in philosophy (epistemology and ethics) and religion (religious semantics in the context of philosophy of science and logical positivism), he is interested in expressions of corporate and organizational culture, as they manifest in the presentation of identity in social media.

1. Introduction

For the past two decades, several models, methods, and strategies have been developed for the purpose of ensuring the longevity of our individual and collective memories in digital form. Archivists and records professionals, librarians, information systems (IS) and IT experts, and various other specialized communities concerned with the preservation of digital artefacts (from space agencies to pharmaceutical companies) have been trying to come up with ‘solutions’ that would ideally be technologically neutral, widely applicable, and cost-effective. Standardization and uniformity are basic ideas that have ever since been guiding those communities’ efforts, as ensuring accessibility and usability of information objects and systems involve some degrees of interoperability and universality. However, it seems to us that the domain-independency goal has been pushed too hard to the detriment of getting an understanding of the specific needs of each environment.

Being prescriptive is another characteristic of traditional approaches to digital preservation—and is a common trait of other areas of intervention concerning the management of records and information as well.¹ This is again a way of moving away or distancing ourselves from ‘what goes on’ in actual situations and focussing on the design of ‘solutions’ that, in their coming close to ‘perfection,’ tend to depart more and more from the real-world problems faced by real organizations and individuals. The models and standards that will be examined in this paper provide the readers with a clear picture of what a digital repository or a trusted records preservation system *should* look like, based on a set of shared assumptions regarding the form records should have if they are to be reliable and authentic, the types of controls that should be in place in order to protect such fundamental record qualities, the actions that those dealing with digital objects are supposed to take, how an organization should overall work to be compliant with professional standards.

As individuals educated in the records management and archival discipline, we do subscribe to the values embedded in the models that we critique; however, we at the same time recognize that as much as the term ‘record’ often means different things to different people, the notion of preservation is anything but neutral or univocal. We argue that a *descriptive* and *situated* approach, an approach that is sensitive to the cultural variations and the ‘centrifugal impulses’ (to borrow from Bakhtin)² that exist between and within any human groups, should become a priority, if digital preservation is to serve the diverse needs existing in society—not just the needs of particular professional communities. By drawing a parallel with theories of linguistics, current approaches to digital preservation seem to focus on the Saussurian ‘langue’ (i.e., “language as a self-sufficient system of signs”) while leaving aside ‘la parole’ (i.e., “speech as the situated realization of the system by particular speakers”).³ Pierre Bourdieu objected to structuralism by saying that “this kind of distinction ... leads linguists to take for granted an object domain which is in fact the product of a complex set of social, historical and political conditions of formations.”⁴ Similarly, all our models and standards only scratch the surface of the object domain of digital preservation. New tools

¹ Fiorella Foscarini, “Understanding Functions: An Organizational Culture Perspective,” *Records Management Journal* 22, no. 1 (2012): 20-36.

² Mikhail M. Bakhtin, “Forms of Time and of the Chronotope in the Novel: Notes Towards a Historical Poetics,” in *The Dialogic Imagination: Four Essays by M. M. Bakhtin*, ed. M. Holquist (Austin, TX: University of Texas Press, 1981), 84–258.

³ Pierre Bourdieu, *Language and Symbolic Power* (Cambridge UK: Polity Press, 1991), 4.

⁴ *Ibid.*, p. 5.

are needed to capture actual instantiations of the processes involved in the creation, management and storage of digital objects and their deep meanings.

Current preservation frameworks take for granted but fail to grasp the specific socio-cultural conditions of records creation and use. In such frameworks, if culture is considered at all, it is mainly understood either as an *abstract* concept or as a *barrier* to the implementation of digital recordkeeping initiatives. The ideal seems to be an organization without culture (i.e., without particularities that make it deviate from the norm). This would make the implementation of systems and strategies straightforward. In some of those studies, knowing the kind of culture you are dealing with amounts to nothing more than using this knowledge to figure out the right way to approach stakeholders in order to push recordkeeping initiatives. We believe that the question is not so much one of labelling cultures as ‘good’ or ‘bad’, or identifying barriers in order to overcome them, because this way of thinking implies that there is one best way of doing recordkeeping that applies to all organizations and oversimplifies the complexity of the situation.

In order to start building the theoretical and methodological framework that would inform what we have named ‘a *culturally sensitive* digital preservation agenda,’ we set out to investigate the ‘information cultures’—that is, “manifestations of organizational culture that portray values and attitudes to information in organizations”⁵—characterizing a number of different semi-public⁶ and private sector’s organizations operating on a global scale and involving different types of industry (e.g., banking and finance, non-profit, information products and services). This innovative framework has been developed primarily by drawing on management and organization theory, IS research, knowledge and information management literature as well as genre theory.

2. Limitations of OAIS and other existing standards and models

The Digital Curation Centre (DCC) in the U.K. defines an Open Archival Information System (OAIS)⁷ as “an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community.”⁸ Thanks to its cross-domain applicability and promises of present and future accessibility and interoperability, the OAIS Reference Model has become *the* standard for building trusted digital repositories. It is a high-level model that specifies the functional entities needed to ensure the continuing preservation of information objects over time in a systematic and structured way.

The OAIS model assumes that the “Producer” shares with the “Archives” the responsibility for—and, before that, an interest in—creating and maintaining meaningful digital assets. The entity Archives takes care of all kinds of information objects *from the point of ingest* and involves a logical, rule-based system for their transformation into preservable, usable, homogeneously structured information packages. On the other hand, the standard does not prescribe or recommend any specific behaviour to be put in place by the Producer, and implicitly suggests that there is no need to investigate the motives that might

⁵ Gillian Oliver, *Organisational Culture for Information Managers* (Oxford: Chandos Publishing, 2011): 4.

⁶ We consider central banks ‘semi-public’ as on the one hand, they are public institutions (not being profit-oriented as commercial banks are) but on the other hand, they do not do business with the general public.

⁷ Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, January 2002. Available at <http://public.ccsds.org/publications/archive/650x0b1.pdf>.

⁸ Digital Curation Centre Glossary of Terms. Available at <http://www.dcc.ac.uk/digital-curation/glossary-terms>.

have prompted the creation of the information objects or any other contextual features ‘in action.’ Negotiations between creating agencies and archives, which would take place any time before the ingest moment and would likely be based on formal descriptions of the agencies’ functions and activities, are considered a sufficiently informative mechanism to feed the descriptive elements involved in an OAIS.

This observation is in line with what Australian archivist Adrian Cunningham wrote in the context of his critique of the idea of digital curation:

The OAIS model makes no attempt to address what is probably the biggest challenge facing digital archivists: How do we find (or indeed ensure the creation of) reliable records that can serve as evidence of decisions and activities among the mountains of what are often dynamic, anarchic, and unmanaged data that organizations and individuals accumulate? If digital archiving is to succeed, it must include intervention in the creation and management of digital information, not just take submission information packages as a given and go from there.⁹

The “intervention” wished for by Cunningham would, in our view, aim at *understanding* what happened in the pre-ingest phase, rather than prescriptively define how the Producer should act in order to facilitate the Archives’ operations.

Not less abstract than the OAIS reference model is the Curation Life Cycle Model developed by the DCC. The DCC model not only refers to a broad category of “data” as its target, but it also covers the whole life cycle of such data.¹⁰ Nevertheless, the model appears unsatisfactory for several reasons. First of all, the series of “Sequential Actions” occurring throughout the life cycle contradicts the notion of simultaneous and recurrent action that the ‘continuum’ model has convincingly put forward since the beginning of the 1990s. Secondly, the model does not elaborate sufficiently on important segments of the life cycle, such as the one labelled “Create or Receive.” Based on which factors, on which elements of the real world, is the organization or individual supposed to “Conceptualize” (i.e., “conceive and plan the creation of data”) and then design and implement a preservation strategy that fits its purpose? Once again, the model appears to be too high level and self referential to be a useful picture of any portion of the reality, to be used as a guiding framework by those who deal with any kinds of data in their daily activities.

More fine-grained and domain-specific are the representations of the chains of actions and relevant products, actors, and mechanisms involved in creating, managing, selecting, and preserving records that were developed by the InterPARES 2 Project and are known as Chain of Preservation Model and Business-driven Recordkeeping Model.¹¹ As recently emphasized by Brothman,¹² these kinds of

⁹ Adrian Cunningham, “Digital Curation/Digital Archiving: A View from the National Archives of Australia,” *The American Archivist* 71(Fall/Winter 2008): 530-543. Citation is at pp. 534-535.

¹⁰ The DCC’s Curation Lifecycle Model is available at <http://www.dcc.ac.uk/resources/curation-lifecycle-model>. In the DCC Model, “data” is defined as “any information in binary digital form.” This includes:

“Digital Objects: simple digital objects (discrete digital items such as text files, image files or sound files, along with their related identifiers and metadata) or complex digital objects (discrete digital objects made by combining a number of other digital objects, such as websites);

Databases: structured collections of records or data stored in a computer system.”

¹¹ Both models are published as Appendix 14 and 15 respectively in the International Research on Permanent Authentic Records in Electronic Systems (InterPARES 2) Project book, available online at <http://www.interpares.org/ip2/book.cfm>.

¹² Brien Brothman, “Designs for Records and Recordkeeping: Visual Presentation in Diplomats, the Record Continuum, and Documentation Strategy,” in *Documenting Society and Institutions. Essays in Honor of Helen Willa Samuels*, ed. Terry Cook (Chicago: Society of American Archivists, 2011), 279–316.

visualizations of the recordkeeping function, which make use of the graphical language typical of engineers with the ill-concealed intention to sound scientific and objective, in reality embed and promote a particular way of approaching the matter, which is rather conservative. Both of the InterPARES models provide a rigorous representation of the ‘right’ way of dealing with records that comes down directly from a noble lineage of archival thinkers, and by doing so, reinforce the values, beliefs, and other cultural manifestations shared by those thinkers. However, the world has much changed since Jenkinson’s times. When tested against most of today’s recordkeeping situations, those abstract models may only work in relation to types of organizations that resemble traditional bureaucracies, while they have difficulty accounting for more dynamic and unstable workplaces.

All current attempts to frame and normalize organizational practices have certainly enriched our understanding of contemporary recordkeeping and have contributed to promote a common terminology and a set of core competences for information professionals, among other benefits.¹³ However, they have also reduced our ability to penetrate and to accept as admissible ‘deviating’ practices, or attitudes towards records and information that only appear to make sense to a relatively small group of practitioners. In other words, those ‘perfect’ models do not help explore actual and therefore complex situations. Additionally, because their design follows some engineering-like approach, those standards are hardly applicable to ill-defined problems, as most of the problems faced by information specialists tend to be.¹⁴

Current preservation concepts appear to have been developed having in mind the features of traditional bureaucratic institutions—also known as ‘full bureaucracies,’ or ‘pyramidal organizations.’¹⁵ Do the cultural assumptions embedded in those configurations still make sense when we examine today’s unstructured, or less structured, and more dynamic and flexible workplaces? Organization theorists have moved beyond the consideration of organizations as monolithic, full bureaucracies, and have identified different social groups within each organization and connected the latter to different ‘sub-cultures.’

The idea that organizations either *have* cultures or *are* cultures emerged as a dominant theme of management literature in the 1980s, when special issues devoted to it appeared in several organization and management research journals. This interest in corporate or organizational culture emerged in connection with a surge of popular interest in new models of managing organizations, and with attempts to explain the success or failure of organizations. In particular, excellence was equated with a ‘good’ or ‘strong’ culture.¹⁶ Soon, the concept of organizational culture had become mainstream, moving into related areas such as IS.¹⁷

¹³ See, for instance, the Records Management standards ISO 15489-1:2001 and ISO 15489-2:2001 issued by the International Organization for Standardization (ISO); the standards relevant to metadata for records ISO 23081-1:2006 and ISO 23081-2:2009; European Commission, *Modular Requirements for Record Systems (MoReq2010)* (2011), available at <http://moreq2010.eu/>; International Council on Archives (ICA), *Principles and Functional Requirements for Records in Electronic Office Environments – Module 3: Guidelines and Functional Requirements for Records in Business Systems* (2008), available at <http://www.ica.org/sites/default/files/ICA-Guidelines-Principles%20and%20Functional%20Requirements%20Module%203.pdf>.

¹⁴ Fiorella Foscari, “Understanding the Context of Records Creation and Use. ‘Hard’ versus ‘Soft’ Approaches to Records Management,” *Archival Science* 4(December 2010): 389-407.

¹⁵ See Geert Hofstede, *Culture’s Consequences. Comparing Values, Behaviors, Institutions, and Organizations across Nations* (2nd ed., Thousand Oaks: Sage Publications, 2001); Henry Mintzberg, *Structure in Fives: Designing Effective Organizations* (Englewood Cliffs: Prentice Hall, 1983).

¹⁶ T.H Peters and R.H Waterman, *In Search of Excellence* (New York: Harper and Row, 1982).

¹⁷ Andrew D. Brown and Ken Starkey, “The Effect of Organizational Culture on Communication and Information,” *Journal of Management Studies* 31, no. 6 (1994): 807-828; Linda Smircich and Marta B. Calas, “Organizational

The majority of approaches in IS research tend to view culture more as a variable property—something an organization ‘possesses’ and which can be objectively measured and manipulated—rather than viewing it as something an organization ‘is’ using in-depth interpretative and longitudinal methods. In the archival discipline (and particularly, in the area of digital preservation), the former appears to be the most popular approach. Studying organizational culture in this context mostly means to identify those attitudes to information that are in conflict with ideal recordkeeping practices, in order to neutralize them, or to minimize their impact, so as to facilitate the implementation of the ‘right’ preservation strategies. Our study considers records and information as socio-historical phenomena whose creation, use, and preservation reflects specific culturally shaped practices, and do not attempt to alter those practices.

3. Organizational culture: A layered view

There is no consensus on the definition of organizational or corporate culture. Edgar Schein’s definition—one of the most quoted—reads: “A pattern of shared basic assumptions invented, discovered, or developed by a given group as it learns to cope with its problems of external adaptation and internal integration, that have worked well enough to be considered valid and therefore, to be taught to new members as the correct way to perceive, think and feel in relation to those problems.”¹⁸ Based on their different visibility to the observer, Schein identifies three distinct levels in organizational cultures:

1. Assumptions (i.e., deeply embedded, taken-for-granted behaviour which is usually unconscious, but constitute the essence of culture);
2. Espoused values (i.e., the organization’s stated values and rules of behaviour, such as, official strategies and philosophies, explicit goals and objectives); and
3. Artefacts and behaviours (i.e., tangible, visible elements in a culture, which can be recognized by people not part of the culture, such as, architecture, furniture, dress code, procedures, records and other “genres of organizational communication”).¹⁹

Since assumptions are generally invisible and pre-conscious, and artefacts, although being more visible, are hard to interpret, most research that attempts to conceptualize culture focuses on “reference group value orientation, such as value dimensions of national culture, or competing values framework at the organizational level.”²⁰

Most recent studies recognize that culture involves several layers, or sub-cultures, and each distinct layer embeds assumptions, values, and practices in different ways. These various cultural layers interact dynamically and permanently to form an individual and an organization’s culture and to shape behaviour. According to Karahanna et al., “depending on the behaviour, different levels of culture will have a

Culture: A Critical Analysis,” in *The Handbook of Organisational Communication*, ed. F. Jablin et al. (Beverly Hills: Sage, 1987), 228-263.

¹⁸ Edgar Schein, *Organizational Culture and Leadership* (San Francisco, CA: Jossey Bass, 1985).

¹⁹ Joanne Yates and Wanda J. Orlikowski “Genres of Organizational Communication: A Structural Approach to Studying Communication and Media,” *Academy of Management Review* 17, no. 2 (1992): 299-326.

²⁰ Dorothy Leidner and Timothy Kayworth, “A Review of Culture in Information Systems Research: Toward a Theory of Information Technology Culture Conflict,” *MIS Quarterly* 30, no. 2 (June 2006): 357-399.

dominant influence on an individual's actions.”²¹ They identified the following types of cultural layers or levels:

- Supranational, i.e., “any cultural differences that cross boundaries or can be seen to exist in more than one nation” (e.g., regional, ethnic, religious, and linguistic sub-cultures);
- National, i.e., “collective properties that are ascribed to the citizens of countries;”
- Professional, i.e., beliefs and practices shaped by education opportunities and acquired professional tenets (e.g., concepts, methods, code of ethics), and that imply a “distinction between loyalty to the employing organization versus loyalty to the industry;”
- Organizational, i.e., “the social and normative glue that holds organizations together;”
- Group, i.e., “cultural differences that are contained within a single group, workgroup, or other collection of individuals at a level less than that of the organization.”²²

The levels of culture are interrelated and can present several configurations. In the case of multinational organizations, organizational culture can span national, religious, ethnic, regional, linguistic, and professional cultures.

Values and practices are two critical components of culture. Values refer to relationships among abstract categories that are characterized by strong affective components and imply preference for a certain type of action. Typically, values are acquired early on in life through the family and neighbourhood, and later through education. They provide us with fundamental assumptions about how things are. Once a value is learned, it becomes integrated in a system of values where each value has a relative priority. Values and practices are intertwined and tend to affect each other. Both are continuously evolving—although values, especially those acquired during the formative years, are hard to change.

Karahanna et al. observe that supranational and national cultural differences are composed primarily of differences in values and, to a lesser extent, of differences in practices. Progressively, the balance between values and practices changes, and practices become a more predominant component of organizational and group cultures. Professional culture implies the acquisition of both values and practices, whereas organizational cultures are made primarily of shared perceptions of organizational practices and to a lesser extent of values. Ideally, practices should reflect values, but this is not always the case. Discontinuity typically occurs when practices dictated by one level of culture (e.g., organization) are at odds with values comprising another level of culture (e.g., professional). An individual whose professional culture is driven by a ‘preservation imperative’ (an archivist, for example) might find it difficult to perform his/her functions in an organization whose practices are detrimental to records preservation. It is not uncommon to witness cultural clashes within task forces or other kinds of ad hoc aggregations, as such groups do often include members from several work units, professions, nations, religions, etc.

As mentioned earlier, behaviour appears to be influenced by different cultural levels according to the task to be accomplished. Thus, practices that involve consideration of values (e.g., archival appraisal) will mainly be influenced by national and supranational cultures; while behaviours that consist primarily

²¹ Elena Karahanna, J. Roberto Evaristo, and Mark Srite, “Levels of Culture and Individual Behavior: An Integrative Perspective,” *Journal of Global Information Management* 13, no. 2 (2005): 7.

²² Karahanna et al., “Levels of Culture and Individual Behavior,” 5.

of practice (e.g., records registration) will more likely be influenced by organizational, and group cultures. Digital preservation is a complex enterprise that comprises both technical and value-based components. It is important to distinguish the various tasks and responsibilities involved, so that the different cultural layers to be considered when planning and implementing digital preservation strategies can be clearly identified.

4. Defining information culture

Information culture is as difficult to define as organizational culture. Most definitions equate information culture with the values, beliefs, behaviours, practices, and attitudes expressing an organization's orientation towards, and use of information.²³ Information culture is inextricably intertwined with organizational culture and involves all the levels mentioned above. Information culture preoccupations emerged in the mid 1980s in the context of the failure of IS projects. These failures were often the expensive result of a misfit between the culture embedded in the technology and that of the adopting organization. It became clear that a technological infrastructure for flow and sharing of information is not enough to ensure success. According to Cabrera et al., successful technology assimilation requires either the technology to *fit* the organizational or subgroup culture, or the culture to be *shaped* to fit the behavioural requirements of the technology.²⁴ In any case, it is clear that values are embedded in both the information created and the associated technology, and that these values are assumed in the underlying work practices that the IT is meant to inculcate.

Most IS research has investigated the cultural factors influencing technology development, implementation, adoption, use, and diffusion. Only relatively few studies have addressed the cultural transformations engendered by the continuous use of a given technology. Leidner and Kayworth suggest that studies need to move beyond trying to use cultural values to *predict* whether or not a group will adopt a new technology to *understanding the dynamics of adoption*. "Culture"—they write—"is less instrumental in predicting whether or not an IT will be adopted than it is in predicting the time of adoption, breadth of diffusion, and the objective of adoption."²⁵ Furthermore, Ortiz de Guinea and Markus suggest that over-emphasis on reasoned action theories by IS scholars has resulted in imperfect understandings of the complexity of human behaviour.²⁶ The implications for preservation are clear. Values, attitudes and behaviours all come into play in a discordant cacophony which challenges the appropriateness of our 'perfect' solutions. We have therefore embarked on a research project to identify the components of information culture to enable the development of a culturally sensitive digital preservation agenda.

²³ Choo, C.W., Furness, C., Paquete, S., van den Berg, H., Detlor, B., Bergeron, P., and Heaton, L., "Working with Information: Information Management and Culture in a Professional Services Organization," *Journal of Information Science* 32, no. 6 (2006): 491-510; Davenport, T.H., *Information Ecology: Mastering the Information and Knowledge Environment* (New York: Oxford University Press, 1997); Marchand, D. "What Is your Company's Information Culture?" *The Financial Post* (23/25 November 1996): 14-15.

²⁴ Cabrera, A., Cabrera, E.F., and Barajas, S., "The Key Role of Organizational Culture in a Multi-System View of Technology-Driven Change," *International Journal of Information Management* 21, no. 3 (2001): 245-261.

²⁵ Leidner and Kayworth, "A Review of Culture in Information Systems Research," 366.

²⁶ Ortiz de Guinea, A. and Markus, M.L., "Why Break the Habit of a Lifetime? Rethinking the Roles of Intention, Habit, and Emotion in Continuing Information Technology Use," *MIS Quarterly* 33, no. 3 (2009): 433-444.

5. Research project description

The research into information culture introduced in this paper aims at increasing our understanding of two fundamental aspects of contemporary, information-driven organizations: 1) their information management approaches—i.e., their ways of using information resources “to answer a question, solve a problem, make a decision, negotiate a position, or make sense of a situation”²⁷; and 2) their information technology ‘appropriation’ modes—i.e., their processes of incorporating technologies (e.g., EDRMS, ECMS) into the organization’s work practices.²⁸ Both aspects have a bearing on the feasibility of preservation, preservation planning and strategies, and more generally, the meaning attached to the word preservation by different human groups. We believe that only through an empirical investigation of the cultural forces at play in the pre-ingest phase, digital preservation may become an applicable concept.

The knowledge of specific work environments gained through the initial stages of this research project—consisting in an analysis of the websites of multinational organizations and a global survey—will enable the planning of subsequent stages. In particular, it will help identify organizations that appear suitable for conducting on-site, ethnographic research. In combining survey and intensive fieldwork methods, this project seeks to overcome the methodological split that characterizes research on information culture. By merging a quantitative study of perceptions of information behaviors and practices²⁹ with an interpretive approach based on ethnographic study,³⁰ this research aims at minimizing the limitations inherent in either approach, thus obtaining a rich and multidimensional understanding of culture.

The choice of international and multinational organizations as a target for this study is related to the objective of moving away from considering national culture values as the main component or determinant of organizational culture, as emphasized in previous research.³¹ Given the trend towards globalization that is typical of any aspects of contemporary society, we thought that studying organizations that are multi-cultural by statute would be relevant to most of today’s workplaces. Additionally, it is a fact that private sector corporations are overall under-researched, especially by the archival and records management community. It is likely that corporate information management in these kinds of organizations, having less legal constraints and being more ‘personalized,’ will reveal attitudes and values to information that might be closer to personal information management types of behaviour, in terms of effectiveness and engagement. That is why we decided to focus on private and semi-public organizations.

The two main components of the initial stages of this research project (i.e., website analysis and survey development) are briefly described below.

²⁷ Choo et al. “Working with Information,” 504.

²⁸ For the notion of appropriation, see Gerardine DeSanctis and Marshall S. Poole, “Capturing the Complexity in Advanced Technology Use: Adaptive Structuration Theory,” *Organization Science* 5, no. 2 (1994): 121-147.

²⁹ See Choo et al. “Working with Information”; Choo, C.W., Bergeron, P., Detlor, B., and Heaton, L., “Information Culture and Information Use: An Exploratory Study of Three Organizations,” *Journal of the American Society for Information Science and Technology* 59, no. 5 (2008): 792-804.

³⁰ Brown, A.B., and Starkey, K., “The Effect of Organisational Culture on Communication and Information,” *Journal of Management Studies* 31, no. 6 (1994): 807-828.

³¹ Hofstede, *Culture’s Consequences*; David Bearman, “Diplomatics, Weberian Bureaucracy, and the Management of Electronic Records in Europe and America,” *The American Archivist* 55 (1992): 168-181.

6. Website analysis: Preliminary findings

This project component consists of a comparative analysis of the websites of several multinational organizations operating in the following industries: information products and services (namely, Google, Apple, and Microsoft), banking and finance (Deutsche Bank, HSBC, and JP Morgan), manufacturing (Tata, Hyundai, and Ford), publishing (Amazon and Springer), and non-profit charitable organizations (The Red Cross [IFRC] and Unicef). The goal of exploring the corporate culture projected by these companies is to get a grasp of their ‘espoused values,’ and to facilitate the selection of a few case study sites within each functional area identified. The questions guiding this study were:

1. What is the self-representation of each company’s organization culture, if stated?
2. What themes distinguish each company’s website, and what might that suggest about their organizational culture?
3. What kinds of information do they seek and provide?
4. What can one infer about the ways in which the companies manage their information and the values they ascribe to it?

Findings from the preliminary analysis of the websites of the information products and services companies and the financial services sector companies provide interesting insight into culture as projected. These findings are summarised below to provide a snapshot of the type of differences that can be detected in superficially similar organisations and that will ultimately play a part in influencing the choices made relating to the ways in which information is regarded and managed.

Google, Apple, and Microsoft, three corporations that are among the giants of software, hardware, and Internet applications and media, portray themselves in very distinct ways. While Google, a relative newcomer having only been founded in 1998 (as opposed to 1976 for Apple and 1975 for Microsoft), appears intentionally to project a corporate culture,³² Apple and Microsoft are much more traditional in the way they present themselves online. Aside from product searches and information about software and services, Apple offers very little corporate information to website users.³³ Microsoft is, in some ways, the midpoint between Google and Apple. Overall, whereas Google wants to project a trustworthy air to consumers, and Apple wants consumers to buy its products, Microsoft seems relatively confident that it is already structuring the technological experiences of most of its website’s visitors. Information about Microsoft’s governance and finances is, for instance, richer and better accessible than that provided by its competitors. Unlike Apple or Google, Microsoft’s primary financial report is presented in HTML format under investor relations, with links to more detailed information and yearly reports, as well as SEC files.³⁴

Three of the world’s largest banks, Deutsche Bank, HSBC, and JP Morgan Chase, have one characteristic in common, i.e., their ‘brand culture,’ which may be interpreted as pride in their history and exhibition of symbolic power. All three companies specifically recognize and speak of their own corporate culture. However, what each lists as constitutive of that culture is not very unique: each emphasizes earnings and stability with respect to finances, conducting business with integrity,

³² <http://www.google.com/diversity/index.html>.

³³ <http://www.apple.com/about/>.

³⁴ <http://www.microsoft.com/investor/CorporateGovernance/Overview/default.aspx>. SEC is the US Securities and Exchange Commission.

environmental sustainability, and social responsibility. All three companies list impressive accomplishments in terms of philanthropic endeavours, and their emphasis on environmental, social, and corporate responsibility is very prominent. Additionally, each reassures their potential and current clients and stakeholders that they are earning profit and well poised to meet any possible global financial turmoil. What seems to distinguish them is a preoccupation with what they think makes their bank the best: Deutsche Bank is convinced that their brand speaks for itself;³⁵ HSBC has spent more time than the others in predicting the future of global markets;³⁶ JP Morgan emphasizes their trustworthiness in that they have the best performing employees.³⁷

A comparative analysis of these two sets of websites will highlight cultural differences related to the distinct goals the industries pursue. However, it is only through further investigations into the actual attitudes towards information of the employees of the companies analysed that the characteristics of national, professional and group sub-cultures can emerge, and we can see beyond the superficial manifestations of organizational culture.

7. Information Culture Survey

This kind of close-up analysis, though still primarily quantitative, will be achieved through the online survey tool that we have developed and are now in the process of pilot testing. The purpose of the survey phase of the project is, again, exploratory, and is going to be complemented by means of in-depth, ethnographic research.

The questionnaire has sections on organizational and societal requirements to manage information (awareness and compliance with regulations and policies, formal/informal decision-making processes, etc.), information sharing, trust in information, and a series of scenarios where values/attitudes/practices are tested (e.g., value of documenting one's work [evidence/memory], belief in the technological fix of every problem, engagement with information-related tasks, feelings towards mobility and online access provided by the new digital technologies—in most scenarios, individual assumptions are confronted with organizational/management assumptions). The objective is to get a sense of the values, beliefs, practices that management and staff of the study organizations associate with information, and their perceptions of the role and effects of information and its management on their daily work and on the organization's operations. We are interested in the views of employees working in all organizational departments and performing a variety of functional roles, not just those in charge of records-related functions.

Data will be analysed by applying primarily, quantitative statistical methods. The measurement of different cultural characteristics and the identification of possible factors responsible for such characteristics will be used to formulate initial hypotheses to be used in subsequent research phases.

8. Conclusion

Existing models and approaches to digital preservation have enabled much progress to be made in developing the necessary technological strategies and systems. However, the messy and difficult factors

³⁵ http://www.deutsche-bank.de/en/content/company/mission_and_brand.htm.

³⁶ <http://www.hsbc.com/1/2/about/advertising>.

³⁷ http://www.jpmorgan.com/pages/jpmorgan/about/culture_new.

that people bring to the mix have only partially been taken into account and the risks involved have been over-simplified. By employing a situated, empirical approach, and by leaving aside any prescriptive purposes, our research into cultural characteristics and relevant factors aims at mapping the digital landscape and identifying the different needs of any stakeholder involved. Understanding and applying the information culture concept will enable the development of a culturally sensitive framework for digital preservation.

The envisaged framework will celebrate diversity and idiosyncrasy of approaches. It will not assume that the 'right' way of doing recordkeeping should be imposed in order to correct 'deviating practices'—which we would rather call 'innovations.' Actual uses of information artefacts, perceived values, and concrete appropriations of relevant technologies will be examined in depth, so that individual, specific, socially and historically situated 'preservation cultures' can be described and used as a basis for action.

Political, Cultural and Professional Challenges for Digitization and Preservation of Government Information in Papua New Guinea

An Overview

Tukul Sepania Walla Kaiku and Vicky Puipui

Abstract

There are political, cultural and professional challenges for the digitization and preservation of government information in Papua New Guinea. First and foremost of these challenges is that digitization and preservation of government information is not a national government agenda. Some other challenges relate to struggles by individual government agencies for a correct definition, understanding, inference, usage and application of the terms digitization, and digitization and preservation. An examination of such challenges should contribute to resulting determination, discussion and a national agenda for digitization and preservation of government information in Papua New Guinea. It is envisaged that the overview serves as a means to attain UNESCO's proposal for the conference to explore the main issues... "affecting the preservation of digital documentary heritage, in order to develop strategies that will contribute to greater protection of digital assets and help to define an implementation methodology that is appropriate for developing countries, in particular."

Authors

Mrs. Tukul Kaiku is a lecturer in Information and Communication Studies at the University of Papua New Guinea. She worked with the Papua New Guinea National Archives and Public Record Services from 1983-1997, the Department of Provincial and Local Government Affairs from 1997 – 2001 and the Public Sector Reform Management Unit of the Department of Prime Minister and NEC from 2002-2004. Her interests are in the areas of Indigenous Knowledge systems and Cultural Documentation.

Ms. Vicky Puipui is acting National Archivist with the National Archives and Public Records Services of Papua New Guinea. She began work with the National Archives in 1993. From 2005-2010 she was in charge of the Lae Branch of the National Archives until recalled to National Archives Headquarters in 2011. She has attended training in Malaysia in Records Management and international conferences in Record and Archives Management.

1. Introduction

In this paper, the immediate issue of contention is that Papua New Guinea as a nation state is struggling in terms of digitization and preservation of government information of cultural and historic significance and heritage. Hence, this overview of political, cultural and professional challenges for digitization and preservation of government information in Papua New Guinea, in which are highlighted inferences, usage and application of the term digitization. Also highlighted are current digitization and preservation projects by selected government agencies as well as references to government information stored in selected cultural institutions. A conclusion will leave off with a summary of the points discussed in the paper whilst a set of recommendations provides a list of suggestions for consideration in view of UNESCOs proposal for the conference.

2. Papua New Guinea

Papua New Guinea is a developing country and nation state located in the Pacific region. It consists of a main island which it shares with Indonesia and outer islands and atolls and is located north of the continent of Australia. The names Papua and New Guinea are respectively credited to European explorers who sailed by and sighted the mainland and islands of Papua New Guinea in the 1500s.

The people of Papua New Guinea are known as Melanesians with skin colour ranging from light brown to dark brown. Prior European sighting, initial contact and colonization, societies coexisted with the natural environment resulting in a state of rich interaction between nature and man. Today, there are more than 800 cultural linguistic groupings each with their own knowledge systems and traditions. Much of the indigenous knowledge systems remain undocumented.

The country went through a chequered history. It was administered by three colonial administrations, Britain, Germany and Australia, went through a war turbulent period during the Pacific War of 1942-1945 and attained political independence in 1975.

At independence in 1975, the country assumed all administrative activities, as legacy of Australian rule, with a Westminster system of government and a national parliament, a Governor General, a Judiciary and an Executive Council and Bureaucracy consisting of government departments and agencies. The country currently has a three tier government system, a national level government and national parliament, twenty-two provincial governments, three hundred and fourteen Local-level Governments (of which there are urban and rural) each of which consists of wards. There are also districts within each province of which there are 89 such. Each of these layers of governments and administrations produce records.

Most government departments in Papua New Guinea create and receive records. In this discussion, we limit the scope of government information to those of significant cultural, social and historic value. These records are held in agencies such as the Papua New Guinea National Museum and Art Gallery and include cultural artefacts of various societies of Papua New Guinea. The Institute of Papua New Guinea Studies holds sound recordings, literature and photographs; documentary heritage of societies of Papua New Guinea. The National Broadcasting Corporation holds sound recordings pertaining to songs and dances and speeches the significant speech being that of the announcement and declaration of Independence by former Governor-General Sir John Guise. The National Archives and Public Records Services of Papua New Guinea holds government records from 1883 for British New Guinea and Papua and post war records of the country. The Papua New Guinea Patrol Report which spans from 1886 and works by government anthropologist, Francis Edgar Williams are among those records of national significance also housed at the National Archives and Public Records Services. The FE Williams papers have in fact been registered as a Memory of the World Collection. The Michael Somare Library at the University of Papua New Guinea holds unique literature materials of the country including the proclamation of British New Guinea on 6 November 1884.

2. Digitization, what is it really all about?

There is a need for a clear definition of the term digitization. In the meantime, inferences, usages and applications of the term digitization are widespread in literature and online discussions. One online source¹ for instance defines digitization as ‘Conversion of analogue information in any form (text,

¹ <http://www.businessdictionary.com/definition/digitization.html#ixzz23yLENnPI>

photographs, voice, etc.) to digital form with suitable electronic devices (such as a scanner or specialized computer chips) so that the information can be processed, stored, and transmitted through digital circuits, equipment, and networks.’ Here the inference is in the application of the term digitization, where computer hardware technology such as flatbed scanners or a slide projector, tripod, and digital camera and software are used to capture analogue information. Images of texts or photographs such as printed photos or taped videos and other images are replicated and converted into an image file such as a bitmap and off loaded either into a computer hard drive or external drives or as JPEG image onto a web page. Hence, in general digitization is the usage and application of computer technology hardware and software in and for the production of digital copies, representations or versions of information held either in texts or symbols and or sound or audio and or photographs and pictorial works or artwork and or film or images and texture maps or such other 3D objects.

The discussion by Roberts² is one which wraps up the concept, inference, usage, application and context of digitization. In his discussion, digitization is an imaging technology used in archives work alongside micrographics technology. It is a recent reformatting and an alternative imaging technology which offers a range of benefits, including integration with other uses of information and communications technology. As an imaging technique, digitization involves a host of activities as Roberts³ discusses.

Digitization involves defining of the purpose and benefits of digitization for its long-term gains including funding and the purpose for digitization, whether for protection and preservation of original records, or to make the records more accessible where these can be easily found for use, and thirdly for security purposes in case of loss.

Records to be digitized for preservation reasons are those which are in poor physical condition and those at risk such as being fragile and those that are frequently requested or used. Those for easy access and use include popular materials such as photographs and other pictorial materials which are appealing and have wide interest, records for inclusion in interpretive products such as online exhibitions etc, discrete groups with a flair of a project. Those for digitizing for security purposes include records with great iconic value or significance to their creators and or archives, those which have high monetary value and or those which are rare.

Specific technical requirements and considerations include technological equipment including hardware and software to use on the one hand, and the objects to be digitized, colour and pixel depth to use and so on are taken on. The question of optical character recognition (OCR) to ensure the digitized copy can be read, searchable, edited and or published with the use of an electronic device, technical specifications are further sought. Also established are benchmarks or the setting of quality standards such as outputs and outcomes to be attained from the project. Planning and management of a digitization project takes on critical questions such as who is to undertake the digitization project, at what cost and where and how.

Those objects and or materials to be digitized must undergo preparations such as, cleaning and repair of materials to be digitized, ensuring materials are in correct order, identification of materials which may require special treatment and types of formats of materials. Even, identification of special equipment such as cradles, preparation of instructions for staff, moving materials safely and securely to digitization location, setting up suitable location and space for materials undergoing digitization,

² David Roberts, “Digitisation & Imaging,” In *Keeping Archives*, 3rd ed., ed. Jackie Bettington, Kim Eberhard, Rowena Loo, and Clive Smith (Canberra: Australian Society of Archivists, ACT, 2008), 402-434, .p. 402.

³ Ibid.,” pp. 402-434.

preparing of targets and rulers to indicate size and clarity of items, establishment of a file-naming structure or protocol for image files from the project. Of course the total number of items must be accounted for to ensure none is missing.

For image capture, questions to take into account are those such as types of equipment for use to regenerate digital copies of other items or records. For instance equipment such as a flatbed scanner where materials are placed face down on a glass plate for scanning and digital cameras. As well, are processing software such as Adobe Photoshop, etc., to absorb the image.

The task of metadata capture involves capturing and maintaining information about the images to provide and assist users to find and use the digital records which were copied, and to support the documentation process for the management of the digitized records, either internally or externally

Considerations in relation to quality assurance ensure quality of imaging and data capture meets the requirements set for a digitization project while management of digitized archives is concerned with ongoing work to ensure special care, special storage requirements and periodic maintenance and monitoring take place. And finally, the output and outcome of digitization ensure availability, access and use of digital records either online and or replicated CD or DVD or paper copies.

Clearly as alluded to above, digitization is a venture that involves a host of activities including careful planning and identification of those analogue materials to be digitized and the type of technological equipment to use as well as the outcome of such a venture.

3. Political Challenges

Given that digitization is a formatting technique which has reached the shores of Papua New Guinea, the following political challenges are highlighted in view of political and administrative arrangements for coordination and management of government information resources at national, provincial, district and local-level government levels.

A cascading political and administrative challenge at national level is the need for management of government information to be a national agenda. Currently it is not. Compounding this challenge is the issue of the lack of a central coordinating agency responsible for overall coordination of government information. This observation is also seen in the guise of a number of agencies of government and stakeholders who are operating in isolation from each other. Collaboration and partnership between government agencies that manage and have an interest in government information is absent. For now the agencies that fall into this category include:

Department of Communication and Information, National Archives and Public Records Services of Papua New Guinea (NAPRS), National Information and Communication Technology Authority (NICTA), Department of Personnel Management (DPM), Department of Provincial and Local Government Affairs (DPLGA), Papua New Guinea National Museum and Art Gallery (PNGNM&AG), National Cultural Commission, National Statistics Office. PNG Electoral Commission, , Information and Communication Studies Strand and the Michael Somare Library both of the University of Papua New Guinea, Department of National Planning and Monitoring, Department of Community Development and the Papua New Guinea UNESCO Office. As well there are twenty two provinces, eighty nine districts and three hundred or so Local-level Governments and six thousand Local-level Government Wards across the country.

In the absence of a coordination mechanism and framework at national level, government agencies have been left to do their own thing. To fill the void vendors and or donor agencies who see a need for

digitization step in. For instance, the digitization of sound recordings of the National Broadcasting Corporation in 2007 by an Australian Aid funded consultant, the Digitization Project of the Mineral Resources Authority of 2008 by the World Bank and more recently, the Digitization project for the Department of Agriculture and Livestock Land Use Files by the Australian National University.

Another political challenge is the need for establishment of necessary information infrastructure, purpose built buildings, network facilities and connections, equipment and staffing at provincial, district, local-level government and local-level government ward levels. Currently, there are twenty-two provinces, eighty-nine districts and at least three hundred and fourteen Local-level Governments. Each of these entities have no purpose built records or archival facilities for the storage and preservation of government information resources.

There is also need for legislation and national policy on digitization and preservation and National Standards on digitization and preservation of government digital information. Currently there are none in place.

The status of the National Archives and Public Records Services is one political and administrative challenge. Currently, the National Archives and Public Records Services of Papua New Guinea is a Branch of the National Library Services of Papua New Guinea with the Office of Libraries, Archives and Literacy (OLA). As well, the post of the National Archivist is below that of the National Librarian and instead is equivalent to that of a Branch Head of the National Library Services. The National Archives needs to be elevated to a level such as that of an Authority or Department with a professionally qualified Archives Administrator and an equally professionally qualified staff capable of formulating and developing policy guidelines and standards for archival practice including digitization and preservation.

4. Cultural Challenges

The following cultural challenges are in view of the concept of imaging and or digitization and preservation of government information of significant national and cultural value for national heritage, research and preservation purposes. This is in view of Roberts discussion above where digitization is an imaging technique for preservation of those records which are in poor physical condition and or those at risk such as being fragile, and those that are frequently requested and or used as well as popular materials such as photographs and other pictorial materials which are appealing and have wide interest and so on.

The Institute of Papua New Guinea Studies out of their own initiative in 2010 embarked on digitizing photographic negative film strips of photographs of people from various areas of the country using HP Standard scanner and a Desktop computer. The intention was to scan the negative film strips with a view to maintain the original quality of the photographs. The scanner setting was set at default mode with the required quality output displayed on the desktop computer screen. The outputs of the scanning procedure were JPEG format photograph copies saved in an external hard disk drive. The originals were in turn stored in archives storage boxes at the Institute of Papua New Guinea Studies mini archives. The staff member who undertook the project was a trained records and archives graduate of the Information and Communication Studies Strand at the University of Papua New Guinea who knew what to do, hence no major hassles were encountered.

Some challenges in this project included imaging limitations where the scanner used was able to only scan about four to five negatives at one instance. The project was able to scan photographic film strips captured with camera equipment that no longer exist. The possibility of viewing the images on film strips was minimal and the scanning project (a reformatting one) made it possible for the images to be

converted to JPEG images which for now can be useful government information resources for researchers and possibly future generations.⁴

The National Broadcasting Corporation, the Voice of Papua New Guinea is a Radio station which is embarking on digitization under the auspices of AusAID, an Australian government Aid Agency. Theirs is a case of enhancing sound recordings by digitizing tapes to computer disks. A graduate of the Information and Communication Studies Strand is the Archivist on location in this organization and an assistant is capturing and compiling metadata for those recordings.

Other cultural challenges include the need for digitization and preservation projects for other government information resources. For instance, the collections of the National Museum and Art Gallery, the Francis Edgar Williams Photographs which have been included in the Memory of the World Register and the PNG Patrol Reports are some government information resources needing to be digitized for enhancing of the images and possibly newer, cleaner and clearer paper copies.

5. Professional Challenges

The following professional challenges for digitization and preservation of government information in Papua New Guinea are in view of the question of what digitization actually involves as presented in Roberts discussion above. These challenges are in relation to the need for a clear understanding and definition of what digitization and preservation are all about and the purposes and other considerations involved in digitization.

One professional challenge for the digitization and preservation of government information in Papua New Guinea is the lack of a Digitization and Preservation division within the National Archives and Public Records Services establishment. Currently, there is none. There is a need for a Digitization and Preservation division with qualified staff to oversee, facilitate and coordinate digitization and preservation of government information projects and programmes across government. The staff will be qualified professional staff in records, archives and digitization and preservation training.

Training is a professional challenge. There is need for training of officers in digitization and preservation practice. Staff in areas requiring digitization and preservation must be those with digitization, records and archival training qualification. This is necessary for proper understanding and identification and systematic control and documentation of those records which will undergo digitization. It was found for instance during the digitization and preservation project of the Department of Agriculture and Livestock that records selected for digitization initially had no file listing for individual folios. A whole book or file would have been scanned without the minute details such as individual folios or page and metadata capture would undoubtedly be a nightmare exercise. The Department did not have a records manager or an archivist. During the digitization project, the contents of each file were not itemized or listed and would have been copied without such an itemized list of such⁵

Another professional challenge for digitization and preservation of government information would be that of the need for a database and inventory of government information resources. There is a need for a survey of government information types and where these are held and which are candidates for digitization and preservation. For instance, there are government information resources of national and cultural significance such as the Proclamation of British New Guinea Protectorate of 6th November 1884

⁴ Robert Natera, questionnaire response, August 20, 2012.

⁵ Vicky Puipui, in an interview July 10, 2012.

held at the University of Papua New Guinea Michael Somare Library. Also there is a tape containing the declaration and announcement of the newly independent state of Papua New Guinea by then Sir John Guise at 12midnight of 16th September 1975. This tape is at the PNG National Broadcasting Corporation. Photographs of the lowering of the Australian flag on evening of 15th September 1975 and raising of the newly independent nation state of Papua New Guinea flag on 16th September 1975 are at the National Archives and Public Records Services of Papua New Guinea. These documents of national historic significance need to be identified and a database to be kept of such..

Digitization and preservation inferences are also professional challenges. Currently, inference, usage and application of digitization and preservation are inadequate and there is a lack of a working definition and understanding of what digitization and preservation entail. A number of cases attest to this:

The first case in point is where digitization and preservation have been assumed to be solutions for easy access and fast retrieval of information. In this case, a certain government agency⁶ with insistence by vendors and donor consultants pulled together active and semi active personnel files from other geographical locations to the national headquarters for the purpose of digitizing such. The idea was to have details of staff in a system in order for personnel information to be easily and readily available in a computer system instead of staff section personnel of the department having to wade through manual paper-based information to search for information. The files were brought to headquarters, lists were not in order, equipment was purchased but there were no qualified persons to conduct the digitization project. The donor agency project also wound down. The files had to be placed in a shipping container for four years and this year had to be relisted and placed into nine hundred standard archives boxes and transferred to the National Archives and Public Records Services. Another agency where digitization has taken place with the intention for information to be in a computer system for easy access and retrieval originally had no lists to the files and records and in the process of digitization, the records were organized into chronological order and scanned/In so doing certain provenance details of the records were disregarded..

Another case in point is in relation to software and hardware vendors. There are cases where vendors are prompting and pushing government agencies into digitization projects. In most instances, the records of some department are not in any systematic order due to poor or nil records management system in place. Currently, one government agency is digitizing its records for the purpose of space saving.

In general, professional challenges here have been limited to the inference, usage and application of digitization and preservation and in most instances, the challenges are in relation to a lack of understanding of the purpose for digitization and what digitization involves. In so doing there is not clear picture and no systematic approach to digitization and preservation

6. Conclusion

Government information in Papua New Guinea comprises records of its historic colonial past and state of nation hood including its cultural documentary heritage and cultural artefacts. Digitization has come to Papua New Guinea, however, there is a need for a definition of digitization for inference, usage and application in Papua New Guinea. Papua New Guinea needs to rise up to the challenges and benefits that can be derived from the application of digitization as an archival reformatting technique and technology. Political, cultural and professional challenges exist in Papua New Guinea for digitization and preservation

⁶ Known to the authors.

of government information. Government agencies need to understand what digitization and digitization and preservation are all about and for government to rise up to the challenges of digitization as a technological invention even if the benefits are those to serve scientific, education and developmental needs of the country and its people.

7. Recommendations

In view of the overview of political, cultural and professional challenges for the digitization and preservation of government information as presented above, the following broad recommendations are proposed for initial attention and action:

1. An agency to coordinate management of government information to be identified by government
2. For a Government Information Task Force or Committee to be established to comprise members of those government agencies and others including the PNG UNESCO Office who have an interest in the area of Information Management and for the Task Force. And for this Government Information Task Force or Committee to coordinate and oversee various aspects of management of government information in Papua New Guinea.
3. For the roles and responsibilities of stakeholders of the Government Information Task Force or Committee including the PNG UNESCO Office to be defined and for the Task Force to take into account some of the challenges discussed in this paper and to build an issues discussed in this paper
4. For the status of the National Archives and Public Records Services of Papua New Guinea status to be upgraded to an Authority Status for it to play an active role as venue for coordination of government agencies in matters of management of government information including digitization and preservation of government information.

References

- Cole, M. Gabriel. "Memories: Preserving the Past with the Technology of the Future." Accessed May 5, 2012. https://docs.google.com/viewer?pid=bl&srcid=ADGEEsGy1Q_jhswG.
- Hedstrom, Margaret. "Digital Preservation: Problems and Prospects," School of Information, University of Michigan. Accessed April 21, 2012. http://www.dl.slis.tsukuba.ac.jp/DLjournal/No_20/1-hedstrom/1-hedstrom.html.
- Maidment, Ewan. "Report on PMB fieldwork in Port Moresby, 15-29 April 2012." Pacific Manuscript Bureau, The Australian National University, Canberra, ACT, Australia, 2012.
- _____. "Overview of the Training Workshop on Digitization and Preservation of Land Use Files, April, 16 to 27, 2012." Pacific Manuscript Bureau, The Australian National University, Canberra, ACT, Australia, 2012.
- Roberts, David. "Digitisation & Imaging." In *Keeping Archives*, 3rd ed., edited by Jackie Bettington, Kim Eberhard, Rowena Loo, and Clive Smith, 402-434. Canberra: Australian Society of Archivists, ACT, 2008.
- Techtarget. "Digitization." Last modified March 11, 2009. Accessed July 23, 2012. <http://whatis.techtarget.com/definition/digitization>.

Current Situation, Problems and Prospects of the Digital Preservation of Documentary Heritage in China

Xincai Wang¹ and Yunxia Nie²

School of Information Management, Wuhan University, China

¹xincaiwang@163.com; ²nieyunxia@sina.com

Abstract

There are a large number of precious historical documentary resources standing for the long history of the Chinese nation, which are of great significance for human civilization. In recent years, the digital preservation of the documentary heritage has made great advances, but there still are many problems. It is argued that the current situation and problem exists in China, which suggests the need for improved documentary heritage preservation in China. Achieving this will require that we constantly improve the laws and regulations, strengthen the research of the theory and practice, pay attention to personnel training, broaden the financing options, expand the digitization category of the documentary heritage, and strengthen the database construction of the documentary heritage preservation in the future.

Authors

Xincai Wang is Professor and Deputy Dean of School of Information Management, Wuhan University, Commissioner of Teaching Guidance Committee of Archival Science of Institution of Higher Education, the Ministry of Education of China, PR China. Research fields include Archival Preservation and Information Management.

Yunxia Nie is a Ph.D. candidate in the School of Information Management, Wuhan University, and a Lecturer at Nanchang University, PR China. Research fields include Archival Preservation and Information Management.

1. Preface

The General Conference of UNESCO (United Nations Educational, Scientific and Cultural Organization) approved the Recommendation for the Protection of Movable Cultural Property (hereafter Recommendation) at its twentieth session on November 28, 1978. As the Recommendation points out, “Except for the immovable cultural heritage, the cultural heritage also includes the documentary forms of movable objects, which means the documentary heritage as the recording and transmission of knowledge and thinking, such as paper archives, photos, films, tapes, machine-readable records and manuscripts, ancient version book, ancient manuscripts, modern books and other publications of special significance.”¹ Based on the recommendations and the efforts of the International Council on Archives (ICA), the Memory of the World List was founded by UNESCO in 1992. Meanwhile, the Memory of the World Programme was launched as a way to preserve the world’s documentary heritage, which includes manuscripts, archives, books and oral history that are deemed to be of such significance and common value to the world’s cultural heritage.

¹ UNESCO, “Recommendation for the Protection of Movable Cultural Property,” http://portal.unesco.org/unesco/ev.php?URL_ID=13137&URL_DO=DO_TOPIC&URL_SECTION=201&reload=1203999345.

“Documentary heritage” is a specific concept in the Memory of the World Programme, but not as a depository institution of a certain type. Chapter 2 in the “General Guidelines to Safeguard Documentary Heritage” (UNESCO, 2002) points out, “It is usual to think of documentary heritage being kept in the museums, archives and libraries, but the Memory of the World is not defined by institutional types or professions. The heritages may exist in a variety of social and communal frameworks with a close relationship between them, which may have an effect on the ongoing survival, safety and use of the heritages.”

Providing advanced preservation strategies for the world’s digital heritage using digital technology is the main focus and task of the Memory of the World Programme. We should make full use of digital storage media to accelerate the documentary preservation process in China, such as disk storage, CD storage, database storage, etc.² Digitizing the country’s documentary heritage is the strongest tool to promote the national culture and to showcase it to the rest of the world. This paper explores the current situation and the problems, and suggests some measures for documentary heritage digitization.

2. The Current Status of Documentary Heritage Preservation in China

The digitization of the world’s heritage was launched by UNESCO in 1992, aiming at preserving permanently the world’s cultural heritage for the use and enjoyment of all. In the 1990s, Europe and America, Japan and other countries had carried out the digitization work of their cultural heritage. For example, the French government made construction of the information network an emphasis, and made full use of the digital resource in teaching and tourism. The America memory programme digitized 5 million documentary items, and provided plenty of information.

In view of the success of digital technology in the field of preservation of foreign documentary heritage, while forced by the huge pressure of the current situation of the preservation of the documentary heritage in China, digital preservation is becoming common practice. As well, the increasing development and popularity of digital technology and network technology has spurred interest in the digital preservation of the documentary heritage in China, which is manifested as follows.

Firstly, the large collection of digitized resources. As the main storage places in China, libraries and archives are good at digitizing the collection. Taking the National Library of China for example, its digital resources amounted to 561.3 TB by the end of 2011, including 466.8 TB of special resources, which accounts for more than 80% of the total collection. And the National Digital Library of China has already been established and is being enriched and perfected constantly. At the same time, the National Digital Archives of China is under steady construction.

Secondly, the digitization of local documentary heritage is emphasized. At present, the National Library system provides a public platform that integrates all parts of cultural information resources in China for all the people in the world to gain access to the local documentary heritage via the National Cultural Information Resources Sharing Project. This project provides online access to a precious documentary heritage that was once secret. For example, the Stele Forest of Beijing Dongyue Temple has already been scanned into the multimedia resource database collected by the Capital Library, and readers can access and research these resources on the library’s LAN.

² Kaifang Tian, “Preservation of the Memory of the World Programme and the Documentary Heritage,” *Information and Documentation Services* 6 (1995): 42. [authors’ translation of original Chinese source]

Thirdly, the digital preservation of minority documentary heritage has been achieved. Since the 1980s, computer operating systems, character recognition systems, font standardisation efforts, etc., have made great progress in the field of the national characters in China. During the years 1997-1999, the international standards of code character sets of Tibetan, Mongolian, the Yi language, Uygur and Kazak were formally introduced into the international code system.³ These achievements are the solid foundations, which make the digital preservation of the minority documentary heritage possible.

There are network columns for the conservation of ancient books in Tibetan and ancient books in Xinjiang at the digital unit of the Conservation Network of Ancient Books constructed by the National Library. And there are many precious minority documents from the three open batches of the List of Archives Documentary Heritage in China, such as the ancient books of Naxi Dongba, the Guizhou “water book”, the Yi Archives, etc., which will be digitized in order to be protected further. As regards laws and regulations, the National Archives formulated and released the interim provisions of the software functional requirements for archives management in 2001, which were the basis and the reference for the functional design and relevant data format for the digitization system of the minority history archives in Yunnan.

Fourthly, there are great achievements in the digital preservation of precious documentary heritage. Through the efficient use of computer and network technology, the most precious documents have been digitized, and mass databases have been established in order to share these invaluable documentary resources. For example, the digital resource of the Conservation Network of Ancient Books has set up a database of the ancient Chinese books, which is accessible worldwide via the Internet, and has also set up navigations for the ancient books bibliographica. And there are network columns of the ancient books, Dunhuang documentary heritage, and the world of the oracle bone in the special database of Chinese Memory. People anywhere in the world can now easily access this documentary heritage at the click of a mouse button. Virtual Reality technology has been used to protect the cultural heritage, such as the 3D digital project of Longmen Grottoes, which is cubic and impressive. During the process of documentary heritage digitization, the precious original documents have been protected, and the information level and the utilization ratio of the documents have been promoted by establishing the network and database, announcing the dynamic information, offering the digital documentary resources, disseminating the knowledge of the documentary heritage, publishing the expert database of preservation, sharing experiences and successful cases, and linking to the world memory web portals.

3. The Problems of China's Documentary Heritage Preservation

As far as it goes, the documentary heritage protection in China has made unprecedented achievements; many precious artefacts have been salvaged and preserved, which has made an important contribution to inheriting and sharing humankind's cultural heritage, and has also aroused concern for the protection of documentary heritage. However, as a coin has two sides, at the same time, the existing problems or inadequacies cannot be ignored.

³ Jian Gong, “Summary of the National Information Resource Researches Recently,” *Modern Intelligence* 10 (2007): 208-211. [authors' translation of original Chinese source]

3.1 Poor preservation of current status

There are abundant and historical documentary heritage artefacts in China. During the of the reform and opening-up, documentary heritage preservation has been effectively strengthened, especially in recent years. With the promotion of the “Memory of the World Programme,” documentary heritage preservation in China has made great progress, such as the Memory of China Archive Documentary List. Nevertheless, the overall situation is not optimistic. Take the archives department with better custody conditions, for example. Some precious historical archives are suffering from varying degrees of damage due to age and other reasons. Some have been endangered, manifested by archive documents that suffer from mildew and rot, boring insects, creases, paper adhesion, yellowing, fading text, brittle and crumbling paper and so on. According to the National Archives’ statistics, by 2005, there were more than 280 million volumes (pieces) in the National Archives. However, it is estimated that there more than 3300 million volumes (pieces) of historical archives formed before 1949. At present, the number of nationwide crucial archives that need salvage and preservation is 1200 million volumes (pieces),⁴ and many of them are considered precious archival heritage. Another example is the precious ancient maps of China’s territory and terrain, with varying degrees of damage amounting to 45.7% of the total, and in which serious damage occurs in 30%. Some ethnic minorities’ archives are endangered or have already been destroyed.⁵

In addition, according to incomplete statistics from libraries, museums and document collection units, Chinese ancient books amount to more than 27 million. In addition to these national collection units, individuals and temples also hold a large number of ancient books. Due to reasons of environmental pollution and human factors, the extent of ancient acidification and embrittlement has accelerated. In the National Library, since the 1960s, due to the increase of acid rain, the situation had been deteriorating. A recent survey shows that the average pH value of Ancient Books in the National Library was reduced to 6-6.5 from 7-7.5 in the 1960s.⁶ As is well-known, once the pH value drops below 5, books become ‘dead bodies,’ and cannot be used again. A large number of precious documentary resources stored by individuals and temples succumb to such phenomena as loss, discarding and burning; in addition, the constraints of custody and environmental conditions make the preservation status quo worrisome.

3.2 Unsound system and rule for preservation

Currently, China has no authority dedicated to documentary heritage preservation standards and management systems, and there is a lack of a national long-term preservation strategy. As well relevant laws and regulations are not perfect. There is only policy to encourage the declaration of documentary heritage, but no detailed reward and punishment system or national policy framework. Compared to foreign efforts, many countries have clear and overall strategies about cultural heritage development and preservation together with science and technology development planning. For example, the “Wealth plan to save the United States” and “National plan of protecting US” proposed by the U.S. government, which aim to stimulate public concern about cultural heritage, strengthen peoples’ regional geographic identity, pride and sense of participation, and thus save the past and protect the future. The French launched the

⁴ Chong Ma, “Research on the Endangered Archival Heritage Preservation Strategies,” dissertation, Beijing: Renmin University of China, 2008.

⁵ Yuejin Tang, “Thought about China Archive Documentary Preservation,” *Archives Science Bulletin* 5 (2007): 31-34. [authors’ translation of original Chinese source]

⁶ Center of National Ancient Books Preservation, <http://news.sina.com.cn/c/2007-05-25/134413075968.shtml>.

“National planning of cultural heritage (technology) research” programme. Funded by the government, this programme involved 53 research groups focused on scientific research on technology, theory, basic research and protection strategies.

On the other hand, the evaluation system is imperfect, and is mainly reflected in assessment indicators that are too extensive. Promulgated by UNESCO, the Guidelines are the implementation of the basic outline of the Memory of the World Programme. China also initiated the China Archive Documentary Heritage Programme in 2000, and then formulated the Selected Rules of China Archive Documentary Heritage. Although the Rules identify seven selection standards—time, region, nation and character, form and style, systematic, and rarity—the assessment indicators are still too extensive. And the Rules only relate to the first-level assessment indicator of documentary importance, such as the “Subject” of indicators reveals that “under normal circumstances, the archives have great value which reflect the development of productive forces, production and technological progress and the major scientific discoveries and invention, and which involve the significant change of relations of production, economic system and political system in various historical periods, and which reflect the typical significance of some major events of a certain stage of development in the history of China in political, economic, military, culture, diplomacy, science and technology, social life.”⁷ Among them, terms such as “significant change”, “typical significance”, and “symbolic” are fuzzy and difficult to grasp in actual meaning. Another example is the “rarity” indicator. “If the documentary heritage is unique, particularly rare,” this implies that it apparently has high value; however, the phrase “particularly rare” is a vague expression and not conducive to specific action. Secondly, the selection field is relatively narrow. China’s documentary heritage is selected by the National Advisory Committee of the China Archive Documentary Heritage Programme. The Committee is composed of well-known academic experts of documents, archives, ancient books, and history, according to the Selected Rules of the China Archive Documentary Heritage Programme. However, of the 113 selected documentary heritages, nearly 90% belong to archival types; non-archival documentary heritages from libraries, museums and other institutions are extremely limited. To a large extent, adding the word “archive” results in public misunderstanding of the documentary heritage list, as it limits the scope of a declaration and leads to the absence of precious documentary heritage that is stored in our libraries, museums and other places, including a large number of precious books or early, classic, block-printed editions. These include manuscripts, such as *Zhaochengjincang*, held in the National Library, *Manichaeism manuscripts of Tang Dynasty*, *Heavenly Creations of Ming Dynasty*, copper typed *Imperial Past and the Present integration of Qing Emperor Yongzheng*, the oldest official set of Confucian books engraved onto stone tablets, the *Xiping Stone Classics*, the silk book “Lao-tzu” of the Han Dynasty, the Art of War Residuals in Linyi Yinqueshan Han Tomb, and so on. The absence of these precious documents creates an obstacle for the further improvement of the documentary heritage list.

According to the UNESCO Guidelines, most of the world’s documentary heritage exists in libraries, archives, museums or other places. However, a lot of the documentary heritage appearing in the Memory of the World List is non-archival and comes from libraries or museums. For example, the “Gutenberg Bible” is Europe’s first lead movable-print; “*BaegunHwasang Chorok Buljo Jikji Simche Yojeol (vol. II)*” is the oldest existing movable metal print in the world.

⁷ XinCai Wang and Wen Li, “Analysis on the Importance Evaluation of the Chinese Literature Heritage,” *Journal of Macau Documentation and Information* (2011) :4-9. [authors’ translation of original Chinese source]

To some extent, as a national list that corresponds to the Memory of the World Programme, the standard of the Memory of China Archive Documentary Heritage List is far from the inclusion standard of the Memory of the World Programme, and thus is insufficient for the digitization of China's documentary heritage.

3.3 Shortage of personnel training

China's personnel training structure for documentary digital preservation is not adequate; the students' knowledge structure is singular, teacher resources are weak and there are temporary shortages. Examining the personnel training of digital preservation for archives, there are fewer than 20 persons in the universities and colleges in China, according to the survey. The teaching content is mainly about traditional paper documents rather than the digitization of archives, which results in students who, after graduation, often are at a loss as to what to do once faced with requests for the digitization of archives.

As a matter of fact, the digitization personnel are not professionally-trained, and they have not received systematic education on documentary digital preservation. In addition to constraints resulting from age, the knowledge structure and the education level, a labour shortage emerges in the field of documentary digital preservation. Furthermore, the work of documentary heritage digitization is monotonous and arduous, and work-place treatment is relatively poor, which often makes the experienced workers eager to leave. Therefore, talent shortages are the main obstacle restricting the fast development of documentary heritage digital preservation in China.

3.4 Weak preservation awareness

In the practical work of documentary heritage preservation, lack of awareness of digital preservation is reflected at both the macro and micro levels. On the macro level, the society as a whole lacks understanding of the documentary heritage, and preservation awareness is weak. Thus, a good environment for protecting the documentary heritage has yet to be created. The micro level refers to the documentary heritage workers who do not follow digital preservation rules strictly, such as ignoring the control of temperature and humidity, and neglecting the impact of polluted air and magnetic interference, all of which are bad for the digital carriers. What is more, some people think that digitization may pose a risk for the safety of information, while others think that it is not necessary to digitize printed documents.

One of the direct effects of weak preservation awareness is to "pay attention to the application, while scorning the protection." For example, although some documentary departments actively apply to be included into the List, once selected, they do nothing more. They believe that once the documentary heritage is selected for the Memory of the World List or the Memory of China Archive Documentary Heritage List, everything is all right, and do not think further preservation is required, such as preservation of the originals or the preservation of digital reproducibility. Moreover, some departments also expose some problems with respect to special funds for digital preservation of documentary heritage. For instance, inappropriate funding, where the funds cannot be earmarked, resulting in the documents needing digital preservation not being digitized. This influences both digital utilization and sharing, and wastes the best time to undertake digital preservation. Just as the famous culture scholar Jicai Feng said, "Applying for a 'World Heritage' is a sacred thing, but now, being in the 'World Heritage' application wave, most government's motives are questionable." Therefore, "regardless of material or non-material

cultural heritages, preservation is more important than the development. If there is no science-based conservation and rational planning, this heritage will ultimately be unworthy of the name.”⁸

3.5 Backward of the large-scale database construction

Although there have been some achievements in the construction of a digital library, digital archives and digital bibliographies of documentary heritage, these are far from the requirements for a special database of documentary heritage that would encompass the whole of the resources shared. Take the documentary digitization of the Republic of China as an example. The National Library undertook the Preservation Project for the Documentary Heritage of the Republic of China, which is a national documentary preservation project for the Preservation Planning of the Ancient Books in China, aimed at screening the documents to remove duplication, while repairing and digitizing them. This helps to avoid the duplication of documents in each place, and also protects the documents of the Republic of China. However, this project does not include the construction of a special database. At present, the database of the Republic of China either is embedded in the digital library project—such as the China-America Digital Academic Library, which is a part of the digital library project invested by the government—or the database is self-established by the institution where the documents are kept—such as the database of the Republic of China developed by the Shanghai Library, the special collection database of the Republic of China developed by Nanjing Normal University Library, and so on.

However, a comprehensive analysis of the present construction of the domestic database reveals that there are many problems, which are described as follows. Regarding content, the database is confined to the library collection, so it is hard to access the total digitized documentary heritage. Regarding the database function, some self-built databases just transform the documents into digital or microfilm format, such as the database of the Republic of China. These databases are not the real database for the documentary heritage. Furthermore, these databases fall short of a uniform standard, which is further compounded by complex problems related to the design structure, the description standard, the retrieval platform, and the browser.⁹ Therefore, it is impossible to make the digitized resources available without preconditions.

3.6 Less capital and backward technology

In China, documentary heritage preservation mainly relies on special funding provided by the central government. Nevertheless, the shortage of funds seriously affects the protection process of numerous documents. For example, although the central budget for the archive documentary heritage preservation fund has increased year by year, because of the huge number of documentary heritage items in the archives in China, combined with too many earlier outstanding accounts, the limited funds are only enough to invest in the salvage of important projects, so the increased funding is like a ‘drop in the bucket.’ Given the lack of supervision, the uneven development of regional economics, the insufficient funding of the archives sector, and the phenomena of misappropriating funds and not dedicating special funds, current conditions for the continued preservation of documentary heritage are extremely unfavourable.

⁸ Jicai Feng, “Concerns for More Protection While Applying the World Heritage,” [authors’ translation of original Chinese source] <http://beijingww.qianlong.com/1470/2010/08/20/229@127376.htm>.

⁹ Qin Sun, “Analysis on the Digital Construction of the Republic of China Records,” *Shandong Library Quarterly* 1 (2008):71-73. [authors’ translation of original Chinese source]

On the technical support level, the level of digital documentary heritage is low, particularly in the selected projects that take fonds as a unit. Due to age and improper conservation, the physical carriers of these documents become fragile. Some archives will break into pieces once touched. Digitization in these instances is very difficult, particularly since the digitization technology mainly depends on scanning, and it is very hard to accurately grasp the repair principle “as the original features.” In addition, the documentary heritage repairs that are undertaken suffer from a lack of matching standards of selection, procedures, methods, and inspection.

4. Prospects on China documentary heritage preservation

In October 2011, the 17th Plenary Meeting of the Communist Party of China put forward the resolution to build a socialist culture authority, so the development of socialist culture will be speeded up. The documentary heritage protection undertakings are bound to usher in greater development in the future to take advantage of this opportunity. Documentary heritage preservation is a huge social engineering undertaking, which not only requires the joint efforts of national, regional and documentary heritage preservation agencies, but also needs the participation, concern and action of every citizen. Given the current status of our documentary heritage protection and the existing problems, there is an urgent need to develop a reasonable and scientific documentary heritage protection system to promote our documentary heritage preservation in an orderly and sustainable manner. Specifically, we should effectively do as follows.

4.1 Improve the legal norms and evaluation system

In the construction of a legal system, we should take relevant international and domestic laws and regulations as reference points to formulate a special law—China Documentary Heritage Preservation Act—or else increase the relevant legal items of the documentary heritage preservation into the existing Cultural Relics Protection Law of the People’s Republic of China, so as to effectively address the problems of documentary heritage preservation and ensure that all of the work undertaken has the necessary legal support. With respect to the development of standards, we should learn from the international Guidelines promulgated by UNESCO and from the selected rules of the United Kingdom, Australia and other countries, and formulate the Selected Rules to China’s Documentary Heritage Programme and China’s Documentary Heritage Preservation Standards. This also should involve refining the assessment of the relevant standard indicators, such as simplifying the second level of evaluation indicators on the basis of the importance of the first level assessment indicators and enhancing the operability of the evaluation indicators, while increasing the percentage of non-archival documentary items, including community documentary items of spiritual value. This will help ensure the preservation of a broader scope and field of distinct documentary heritage in China by helping avoid the underreporting of items, and paying equal attention to both local fields and international fields, thus improving the overall degree of standardization of China’s documentary heritage preservation.

4.2 Deepen theoretical research and strengthen resource integration

Theoretical research is the summary and promotion of the digital preservation of the documentary heritage. In the future, the theoretical research should pay more attention to the development and

promotion of feasible measures of digital preservation and offer practical suggestions for carrying out such work.

Firstly, we should conduct multi-interdisciplinary research into the damage mechanisms, protection environments and preservation methods of the digital heritage carriers, absorbing the newest achievements of modern materials science, to establish the basic theories, concepts and procedures of great general significance for providing the best practical guidance for documentary heritage digitization.

Secondly, we should study and formulate a list of documentary preservation that is inclusive and cohesive. If the list is established, it could rectify the awkwardness of the current “archives only” nature of the list. As well, it could help avoid the problems of ‘list fighting,’ resources distribution, and the division of policies from various sources. It is suggested that we integrate the Memory of China Archive Documentary List and the National Precious Ancient Books List, and establish the Memory of China Documentary Heritage List, while concentrating the necessary manpower, material and financial resources to strengthen the joint application of documentary heritage preservation in order to avoid preservation rivalries among the various documentary heritage sectors.

Finally, we should integrate research resources. On the one hand, researchers of the digital preservation of the documentary heritage should work together to ensure effective communication, dialogue and cooperation, thus creating a ‘big stage’ for the digital preservation of our documentary heritage. On the other hand, the archives, libraries, and museums should keep developing features and advantages among themselves, as this is helpful to ensure a diverse culture.

4.3 Strengthen public awareness and personnel training cultivation

In the future, our country should take advantage of the advanced international experience to preserve the documentary heritage as well as enhance public awareness. In particular, we should actively and publicly promote the importance of documentary heritage preservation in all sectors of society. This could include, for example, carrying out interactive preservation activities during “World Cultural Heritage Day” to improve social awareness of the issues and to promote a collective consciousness about documentary heritage preservation that will, in turn, engender a good preservation environment. Preservation of documentary heritage cannot be separated from the personnel training, so shortage of professional preservation and management personnel is a bottleneck restricting the development of documentary heritage preservation. To solve this problem, on the one hand, we should strengthen the introduction and exchange of international talent; on the other hand, we should encourage the self-cultivation of professionals, gradually form the talent echelon for archives preservation and research, and enhance further the national self-consciousness of cultural inheritance.

4.4 Broaden the financing path and manage funds strictly

The current financial allocation to documentary heritage preservation is the most important and stable funding source, but it often is too little to fully carry out the work, so we should broaden our funding base for documentary heritage preservation. For example, we could actively encourage private capital by outsourcing documentary heritage preservation to businesses, and so on. In this area, we can learn from

the successful documentary heritage preservation experiences of the USA's "the founding era",¹⁰ to engage actively in private fund-raising activities while trying to secure more national funding. In particular, we should strictly implement a policy of "a fixed sum is for a fixed purpose" for the allocated documentary heritage preservation funds, and increase both social supervision and the punishments for the misappropriation of funds. We should implement a project schedule for the process of special documentary heritage preservation, implement a mid-term project inspection measured against acceptable project development, comply with the documentary preservation and repair principles, and put in place strict quality controls. Furthermore, we should ensure that we do not delay the timing of these actions so that we do not miss the best preservation opportunities.

4.5 Innovate the technology of preservation and accelerate the construction of the database

Because advances in digital technology happen quickly, and because the preservation objects are complicated, documentary heritage digital preservation is a dynamic process. Consequently, digital scanning technology based on optical character recognition (OCR) technology alone cannot satisfy the preservation requirements. Instead, we must continually explore new digital technology to meet the different needs, and strive for the most cost-effective approaches that balance speed and quality during the digital preservation of documentary heritage.

At present, the level of digitization of the precious documentary heritage is not high in China, and the database is still at the initial stage. In the future, we should pay more attention to the database of the local, minority and the special documentary heritages. In particular, we should speed up the pace of digitization of the original documentary heritage items that are not included in the Memory of China Documentary Heritage List, construct a platform of preservation and management, study and establish the huge database mode for the documentary heritage preservation, establish a comprehensive database management system, and improve the capacity of the system against risks. In addition, we should emphasize the construction of standard and special collections, while importing advanced databases and building the special databases ourselves. Only in this way will we be able to construct a digital preservation system for our documentary heritage.

4.6 Promote the digitalization and cover all kinds of documentaries

China's documentary heritage is wide ranging, and the standards of classification vary widely. Depending on the locations of individual documentary heritage items, they could be preserved in public institutions (archives, libraries, museums, etc.), or be kept with folk organizations or in the custody of individuals. The different types of documentary heritage items include minority documentaries, Republic of China documentaries, the Cultural Revolution documentaries, celebrity documentaries, and so on. At present, the objects of digital preservation are the documentaries kept in the public literature institutions—i.e., the local documentaries, the minority documentaries, and the Republic of China documentaries. However, the Cultural Revolution documentaries, celebrity documentaries, and scientific and technical documentaries, which are rich in social, cultural, historical and economic values are not included in the digital preservation planning. In the future, we should gradually expand and extend the scope and depth of the

¹⁰ Jianghua Wu, "Documentary Heritage Preservation Process in the USA 'the founding era'," *Archives Science Bulletin* 4 (2011): 30-33. [authors' translation of original Chinese source]

digital preservation documentary heritage to include these other areas so that our entire documentary heritage is covered.

5. Conclusion

At present, the documentary heritage preservation programme has begun to take shape and has achieved some results in China. However, there are a lot of problems at the same time. The future road of documentary heritage preservation is long and tortuous. Only if we are able to take full advantage of the rich documentary heritage resources in China and effectively solve the relevant problems will be able to make our valuable documentary heritage available to the Memory of the World List and allow the rest of the world to share more fully in China's cultural achievements.

Additional References

1. Yaolin Zhou. *Theory and Practice for Archive Documentary Heritage Preservation*. Wuhan: Wuhan University Press, 2008.
2. Jiazhen Liu. *Documentary Heritage Preservation*. Higher Education Press, 2005.
4. Guoqiang Wang. "Summary on the Realistic Value about the Preservation Methods of Ancient Books." *Library and Information* 1 (2010): 30-34.
5. Lizhu Guo. "Fundamental Countermeasures of Archives Heritages Preservation in China." *Archives Science Bulletin* 3 (2006): 57-59.
6. Quan Zheng. "Research on the Building Mode on Ethnic Minorities Documentary Heritage Preservation." *Archives Science Bulletin* 3 (2009): 67-71.
7. Yuanming Peng. "Research on the Methodology of Archive Documentary Heritages Preservation and Utilization," dissertation, Shanghai: Fudan University, 2008.
12. "Memory of China Archive Documentary List," <http://baike.baidu.com/view/3477472.htm>.
13. "7 Items are Selected in Memory of the World List," <http://www.csstoday.net/Item.aspx?id=7227>.
14. "Memory of the World List," <http://baike.baidu.com/view/1010083.htm>.
15. "Scientific Management Strategies of Repairing and Salvaging for Key Archives Programme is Agreed to Review," <http://www.saac.gov.cn/articleaction.do?method=view&id=ff8080812a944627012ccdd71328008a>.
16. "National Precious Ancient Books List," <http://baike.baidu.com/view/2362437.htm>.

Open Archival Information System Reference Model: Answer or Inspiration?

Digital Archiving Systems Confronted with the OAIS Reference Model

Stefano S. Cavaglieri

Abstract

The OAIS reference model defines Archival Storage as the entity that contains the services and functions used for the storage and retrieval of Archival Information Packages. Can we use this term to define a digital archiving system? A clear definition of Information is central to the ability of an archiving system to preserve it. A system can be said to have a knowledge base, which allows it to understand received information, where information is defined as any type of knowledge that can be exchanged, and this information is always represented by some type of data. In the digital domain, in order for bits and bytes to be successfully preserved, it is critical for an archive to clearly identify and understand the data and its associated representation. This requires transparency to the bit level, and is a distinguishing feature of digital information preservation, which runs counter to trendy object-oriented concepts. This presents a significant challenge to the preservation of digital information.

Author

Stefano Sergio Cavaglieri is Chief Technology and Information Officer for the Swiss National Sound Archives in Lugano, Switzerland. His career started back in the late 70s in the audio engineering field. In the late 80s, in parallel to audio, Stefano broadens his interests with computer science. His expertise and analytical skills lead him to be now one of the most recognized technical authorities in the multimedia archives community. Stefano holds a degree both in electroacoustics and in computer science. He serves as an active member of IASA – TC, ALA, and AES.

The Open Archival Information System (OAIS) reference model, a standard (ISO 14721:2012) now widely adopted all over the world by all kind of archival organizations, defines Archival Storage as “the entity that contains the services and functions used for the storage and retrieval of Archival Information Packages (AIP).”

The terms “archival storage” and “digital archiving system” are often confused with each other and mistaken, probably because both contain the word “archive”. But archival storage is only one of the various components that constitute the OAIS functional model, and only one piece of a proper digital archiving system, which should be understood as a bunch of physical, as well as logical parts, all aimed to the preservation of digital information. The basic set of functional entities of such a system is made up of Ingest, Archival Storage, and Data Management. While the full set expands it by adding Administration, Preservation Planning, and Access. There is no one-to-one correspondence between each single functional entity of the OAIS reference model and any physical or logical component, the only exception being perhaps any non-digital parts, such as a printed manual for instance.

A clear definition of Information is central to the ability of an archiving system to preserve it. A system can be said to have a knowledge base, which allows it to understand received information. Information is defined as any type of knowledge that can be exchanged, and this information is always represented by some type of data.

As an example, the information in a printed book is typically expressed by the observable characters which, when they are combined with a knowledge of the language used, are converted to more meaningful information. If the recipient does not already include this language in its knowledge base,

then the written text needs to be accompanied by relevant dictionary and grammar information, in a form that is understandable using the recipient's knowledge base.

The same should apply in the digital domain. In order for bits and bytes to be successfully preserved, it is of essential importance for an archival organization to clearly identify and understand the data and its associated representation. This requires true transparency to the bit level, and is a distinguishing feature of digital information preservation, even though it runs counter to the object-oriented concept and trend, which indeed tries to blur everything behind a user interface. This presents a significant challenge to the preservation of digital information.

The software layer that is typically used to access digital information, does incorporate some understanding of the network of representation information objects involved. However, the fact that we have this piece of software at our fingertips should not be used as rationale for avoiding identifying and gathering readily understandable representation information that defines the information object. Please keep in mind that it is harder to preserve working software than to preserve information in digital or hard-copy forms.

Having said that, when designing and implementing a digital storage system conforming to the OAIS reference model, we are requested to think and act with absolute coherence from the top to the bottom. We must ensure that the information to be preserved is independently understandable to the community. In other words, a series of bits, usually packaged into a file, should be accessible transparently, without needing any third party assistance, all the way long, down to the physical storage media, being it a disk, a tape, a chip, or whatever built into the digital archiving system.

A key to success, on this matter, for organizations such as libraries and archives, is to consider a digital archiving system in a very much similar way as a traditional, physical archive, and to pretend it to function likewise. But how? Let's start by tracking a typical journey of a physical sound carrier, from the very moment it enters the archival premises, down to its final storage in the vaults, keeping constant, side-to-side comparison with its digital derivative:

- A sound recording is delivered by means of a physical carrier (i.e., a CD, a vinyl disc, a tape, etc.) to the archival organization. There, it gets visually inspected for damages and, if no visible issues are detected, its acquisition, or entry, gets recorded. Otherwise, it is sent back to the distributor, asking for a replacement copy. If this is the case, the process restarts from the beginning. In OAIS terms, this corresponds to the interaction between Producer and Archives, by means of a Submission Information Package (SIP), under Administration control.
- Next step is to proceed to inventorying, where our sound carrier is assigned a call number, which identifies the specific Item we hold in our hands. This is the beginning of the Ingest process, where we start generating an AIP.

Already at this very early stage, if the item (i.e., the sound carrier) is solely identified by a code only readable by a machine, such as a numeric barcode for instance, its accessibility starts relying on the availability of the logistics management system, which says number "n" belongs to room/rack/shelf "xyz". Take off, for any reasons, this set of hardware machinery and related software, and you're done! In case your organization spans a number of vaults, storing millions of items, your specific sound carrier might be buried forever, and its preservation compromised as a consequence.

You can prevent this from happening, and make life easier for your archive staff, by defining and strictly enforcing a clear naming convention. Our recommendation: the call number assigned to an item should be some sort of a human-readable, and understandable, unique identification, possibly relating to

the item's physical location within the storage facility. It shall contain enough information for an archivist to understand what room, what rack, what shelf to look for, first to archive it and later to find it again whenever needed.

In OAIS terms, the naming convention defined for-/ and applied to a call number, together with some basic understanding of the logistics of the vaults, constitute the knowledge base, which allows the archivist to understand the received information.

- Cataloging comes next, where unstructured data is read from the sound carrier itself and its container, which may be an envelope, a box, or any annex. All data is then structured (i.e., split and normalized) according to some standard set of cataloging rules. Assuming we adopted the Resource Description and Access (RDA) standard (as an OAIS note—Did we already include this information in the overall knowledge base of our organization?), which embeds the Functional Requirements for Bibliographic Data (FRBR) model, all document-wide data is grouped and entered as a Manifestation entity, while each track-related set of data is entered as an Expression entity, and refers respectively to a Work entity.

If we, as many of us already do, use a database system for keeping track of the information, each of the above entities usually gets its own identification (ID). Most of the times this IDs are automatically generated and assigned, maybe hidden, by the system, then cross-linked in a way or another. Is the linking mechanism transparent enough? You may ask your vendor whether it is possible or not to determine the relationship between the various documentation entities (i.e., talking databases, between the various records and files, or rows and tables, depending on the underlying architecture) in case we'd take off this particular piece of software. When not, your collection is in danger, although a possible, but not always viable workaround would be to enter some additional linking information "that makes sense" into each and every record. Here again, in OAIS terms, we are in a situation where our best preservation intents might get locked in, in this case by some technology.

The descriptive information (i.e., the various cataloging entities) is an additional sensitive point by itself. Does it only exist as entries in a database? Is the filing system of the database proprietary? Is the data encrypted somehow? Did anybody decide for any particular character sets? Those are some of the questions an archival organization aimed to be OAIS compliant, should ask itself. As yet another recommendation: if one or more answers to the above questions are positive, we better think of protecting ourselves by storing this descriptive information in parallel (i.e., a periodic export from the database system would do the job), at the operating system level, as text-only, such as XML-based files for instance.

At the cataloging stage, the sound recording we received at the very beginning of this journey by means of a sound carrier, gets to life within the digital archiving system (i.e., the electronic catalog itself is a system component)—now its avatar exists and can be further exploited.

- Next, the sound carrier is taken to the re-recording facility, where it gets "digitized". In case of an analogue sound recording, its content is extracted using a suitable player and converted from analogue to digital (i.e., we generate a sequence of numbers, representing the amplitude of the original signal at specific points in time). In case of a digital sound recording, its content is simply transferred (i.e., the numbers are simply extracted as-they-are). This series of numbers (i.e., the digital information, also known as "essence") is then packed and stored as one or multiple entities into a particular filing system, which may consist of records, when talking Document Management Systems (DMS), or files, when talking straight file-/ operating systems (FS).

It would make sense to keep some sort of a relationship between the entities just generated and the original sound carrier. What about reusing, extending when necessary, the same naming convention already in place for assigning the call number? As of today, this is generally possible in a FS environment. It is true that file systems enforce some limits concerning the use of several specific signs, they may also make a difference between upper-/ and lower case, but as long as we are concerned with numbers, plain Latin alphabet characters, and some basic punctuation, there is no particular restriction. Assuming we agree with the initial statement and question, why don't we take our notion of room, rack, and shelf one step further and also apply it to the file system architecture (i.e., the directory structure, for instance)?

As an example, let's take a Compact Disc (CD): according to our naming convention, we assign it a call number, such as "CD123"—where "CD" refers to the format, thus to the vault and rack type (assuming we have some dedicated, optimized space for CDs), and "123" is a sequential number that, together with some range labels on the racks, refers to a position on a specific shelf. Now take its digital derivative: the contents extracted are a series of files, sharing the same call number, identified such as "CD123_1.wav"—where "CD123" refers to a (sub)directory in our file system, "_1" refers to track #1 on the original CD, and ".wav" refers to the file format we are actually storing. If all data we derive or extract (sound, images, text, etc.) from the original CD is treated and stored the same way as for the above file, following the same principles and rules, and put into the same directory, this directory potentially becomes our AIP, thus fully OAIS compliant.

In a DMS environment, things tend to get more complicated, in a very similar fashion as earlier described for the linking of cataloging entities. We can re-raise the cross-linking issue by saying we need some sort of a human-understandable mapping between the IDs assigned to these "essence-records" by the system and the real world. In case of a negative or non-exhaustive answer from the system vendor, (we're still trying to conform with OAIS, right?) we are once again in danger of getting locked in by some technology.

- As a last step, our sound carrier is prepared for long-term archiving (i.e., it is cleaned and repackaged, if necessary), then taken into a secured, climate-controlled vault, and placed in the rack and on the shelf it belongs to, according to the information embedded in its call number. From now on, only its digital derivative is delivered to any patrons or customers, on demand, as a Dissemination Information Package (DIP).

We may think we are on the safe side now—we just put a number of potential issues on the table, and for some we even found a solution—there comes the next problem: the storage. Digital storage in not equal digital storage, under several perspectives. Let's start from the core, the storage medium:

- Is it a hard disk? Besides any sudden crashes, as a very common head crash for instance, the data stored on it looks quite safe... as long as you keep it spinning—otherwise strange behavior as bit-flip may occur; and attached to the same device—its input/output bus might become obsolete; and you drive it with the same operating and file systems—although there's a certain cross-compatibility between those, no guarantee you won't lose data when you switch it.
- Is it a disk array? The security level is somewhat higher, compared to a standalone disk, and very much depends on the so called RAID level we adopt. Assuming we have enough resources, and go for a most reliable level, as RAID-6 for instance, which can tolerate the concurrent failure of no less than two hard drives, are we aware that data, when written to, is automatically split across

several disks? And that just removing one of such disks from the array, trying to read it as a standalone unit, will turn into a miserable failure?

- Is it a cartridge, perhaps in a robotic library? The security level becomes even higher... provided that, among other things, the data written on the cartridge are directly accessible, thus unmodified by any software layer sitting in front of it; the drive needed for reading this specific format is also available as a standalone device and it can connect to your computer.

If you are lucky enough to own-/ or plan for one or more cartridge-based robotic libraries, we strongly recommend to ask your vendor (i.e., you better do the test personally) whether it is possible or not to do the following with your system: write a file to the system using its full functionality; take note of the ID of the cartridge your file is being written to; brutally switch-off the system, open the door of the robotic library and pull the affected cartridge; connect a suitable drive to your computer (please consider running some version of Unix or Linux) and insert the cartridge; mount it as a regular volume and eventually try to find and to read your file. Was your test successful? If so, you are one step closer to OAIS compliance. If not, you may, sooner or later, get into big troubles.

Although not exhaustive, this list of considerations somewhat scratches the surface of the subject as in the title of this paper. The main intention is to give a basic understanding of it, to highlight a number of difficulties and potential points of failure, some of it very basic, as well as to show the level of complexity everyone is facing when trying to design, to implement, and to eventually run a proper OAIS conforming digital archiving system.

One might even wonder, at this point, whether it is possible or not to get through and to solve all these issues. The good news is: this is possible, although rather expensive, in terms of resources required! The bad news is: it is fairly supported by most solution providers. Why are they not supporting it?

Anyone familiar with the words benefits and revenues? And the fear of seeing any glitches on the mid-/ to long run? This is probably also due to the difference in perspectives, when talking long run—for a typical Information Technology (IT) company “long run” is something between 5 to 10 years, while for an archival organization it means just forever.

Talking resources, money is unfortunately the foundation for keeping digital archiving systems alive. Need some arguments? Each and every digital archiving system relies on energy—mainly for keeping it running and cooling it down—and the amount of energy required can be huge, depending on the system architecture, the place it is located, etc. Handling and management of such systems need very specific skills—which translates in expensive training for resident staff and maintenance contracts—and dedicated staff, in case of medium to large scale systems. IT components in general become obsolete very quickly—hard disks have a typical lifespan of about 5 years, cartridges would last longer but new generations are introduced approximately every 3 years, not to mention any software related issues—thus major migrations must be regularly scheduled.

All this means you are required a constant flow of money, from the very first day you even think of your own digital archiving system until a new revolutionary, universal archiving technique is invented. This may sound frustrating but it sums up the current, real situation.

What to do if you are a smaller organization? Think of being in a privileged situation. In fact, an attractive, complementary option to what discussed so far is merging (i.e., harvesting) your collections and problems with a conforming OAIS archiving system of a larger organization. This has lots of benefits and almost no negative side effects.

In conclusion, the main advantage offered by a digital archiving system is that it is self-contained, which means everything is available from the same source, at your fingertips. In some cases, as in the audiovisual world for instance, it is even the sole means for preserving (at least part of) our cultural heritage. While the main disadvantage is that it is made up of a large number of unstable components, which are assumed to keep talking to each other. With the term “unstable” I mean that each and every component is constantly updated, in some cases enhanced or even replaced with new ones, which requires the perpetual allocation of enormous resources.

Besides all that, beyond technology, the most important thing we should never forget is that the knowledge base is your life-jacket—everything (procedures, standards, workflows, technical insights, etc.) should be documented, to the bone... preferably on plain paper.

A Model for Managing Digital Pictures of the National Archives of Iran

Based on the Open Archival Information System Reference Model

Saeed Rezaei Sharifabadi, Mansour Tajdaran and Zohreh Rasouli

Alzahra University, Iran

Abstract

The present research is to represent a model for managing digital pictures of National Archives of Iran, based on the Open Archival Information System reference model (OAIS). The research is quantitative and uses the library method. Descriptive and analytic survey is also used for data analysis. The data are gathered through a researcher made check list. The statistic population involves 32 persons, expert at digital preservation field and familiar with OAIS reference model, among whom only 20 persons accepted to complete the check list. Regarding experts' viewpoint, 89 elements from the total 135 elements available at audit list, have been represented for the suggested model of managing digital pictures of National Archives of Iran. Digital pictures are important records which have considerable role on the meeting of users' information needs. Since there has been done no research at the field of managing digital pictures of National Archives of Iran, based on OAIS reference model so far; this research is to represent a desirable model for managing digital pictures of National Archives of Iran through developing a check list based on OAIS reference model.

Authors

Saeed Rezaei Sharifabadi received his Ph.D. degree in Library and Information Science from the University of New South Wales (Australia) in 1996 and is currently an Associate Professor and the Dean of Faculty of Education and Psychology at Alzahra University (Iran). He has served as the Editor-in-Chief of *Ganjine_ye Asnad* (a historical research and archival studies quarterly) since 2006. Dr. Rezaei is also former Deputy National Archivist of Iran and a former member of the Iranian National Committee of the Memory of the World Program. He has been part of the InterPARES Project during his sabbatical leave at the University of British Columbia in 2009. His publications/research deal mainly with IT Applications in Libraries and Archives, srezaei@alzahra.ac.ir.

Mansour Tajdaran is an Assistant Professor in the Department of Library and Information Science at Alzahra University, tajdaranmansour@yahoo.com.

Zohreh Rasouli holds an M.A. in Library and Information Science from Alzahra University, rasoulizohreh@yahoo.com.

1. Introduction

Today, many information centers have digitize their resources. With the development of technology, various formats developed the old softwares and hardwares are replaced with the new ones. All these cause a big challenge for the long-term access to the content of digital resources. Therefore, applying the digital life cycle, which is responsible for the preservation and long-term access, is considerably important. Many national libraries and archives have benefited from OAIS reference model for designing and constructing their digital archives. Currently, this reference model is a well-known standard in the field of digital preservation that has been approved by ISO in 2003. The OAIS has created a general

model or framework for the construction and maintenance of information warehouse for the long-term preservation and access to digital materials (Samiee, et al. 2011, p. 165).

2. Research Questions

The present research is an attempt to answer the following research questions regarding the experts' viewpoints:

1. What are the most important components of the OAIS reference model ingest functional entity for managing digital pictures of the National Archives of Iran?
2. What are the most important components of the OAIS reference model archival storage functional entity for managing digital pictures of the National Archives of Iran?
3. What are the most important elements of the OAIS reference model data management functional entity for managing digital pictures of the National Archives of Iran?
4. What are the most important components of the OAIS reference model administration functional entity for managing digital pictures of the National Archives of Iran?
5. What are the most important components of the OAIS reference model preservation planning for managing digital pictures of the National Archives of Iran?
6. What are the most important components of the OAIS reference model access for managing digital pictures of the National Archives of Iran?

3. Review of the Related Literature

The review of the related literature which includes the previous studies conducted inside and outside Iran are further discussed in details.

Samiee (2010), in her Ph.D. dissertation titled "an Examination of the digital preservation posture of the resources of national libraries members of the Internet Preservation International Consortium and representing a suggested model for the National Library and Archives of Iran," studied and analysed the OAIS reference model as the only digital preservation standard in digital repositories and suggested the implementation of this standard at digital libraries' software.

Beedham et al. (2004) in a research paper titled "Assessment of Ukda and Tena conformity rate to METS (Metadata Encoding and Transmission Standard) and OAIS reference model," examined six main entities of the OAIS reference model and METS and then explained about their functions at digital archives such as Tena and Ukda.

Sample (2004) in a research paper titled "Developing strategies for Digital preservation in Edinburg Academic Library," studied the general view toward developing preliminary project of digital preservation in Edinburg University; the integrity of METS; the OAIS reference model and development of digital preservation operations at digital materials system in Edinburg Academic Library.

Anderson et al. (2006) in a research paper titled "An archival study of digital photos," described digital photos features, preservation strategies, metadata available for digital photos and digital photos life cycle at the OAIS reference model and trusted digital repositories (T.D.R.).

Nordland (2007) in his Ph.D. dissertation titled as “International Development and Research Center (IDRC), a case study for long-term preservation strategy, has considered challenges for archivists and records managers at long-term preservation management of digital records. This dissertation has studied IDRC as a prominent instance in different areas of digital long-term preservation such as digital preservation standards and strategies, permanent digital repository, web archiving and alike. The results of this research have contributed to the OAIS reference model for maintenance and preservation of information resources at archives.

The reviewed studies indicate that the OAIS reference model is a formal standard at digital preservation area and using it at digital repositories will contribute to the integral management of existing digital resources, saving cost and time for setting up digitizing systems and easy sharing of and access to information. This research aims at suggesting a framework and model for managing digital pictures of the National Archives of Iran, based on this model (OAIS).

4. Research Methodology

Regarding the content of this research, it is a survey study (descriptive and analytic). First, OAIS reference model was studied in order to determine the prominent activities at main functional entities of reference model, including: ingest, archival storage, data management, administration, preservation planning and access. Final entity is common services which underlies all the others, and includes operating system services, network services and security services.

After determining these activities for data gathering, a list of 135 closed questions—based on Likert scale (5 options)—was designed which included the 6 main entities: ingest (23 questions), archival storage (21 questions), data management (21 questions), administration (28 questions), preservation Planning (25 questions) and access (16 questions). Content validity method has been used to assess the researcher made assessment list validity. After preparing the checklist, experts in the archival field and members of scientific board were asked to assess its validity, and then the suggested revisions were made, Cronbach alpha coefficient was used for assessment of the checklist reliability at this research; the checklist reliability has been calculated through Cronbach alpha coefficient, separated into 6 main entities of the OAIS model (Table 1).

Table 1: Cronbach Alfa quantities, separated into 6 main entities of OAIS model

number	OAIS reference model entities	Cronbach Alfa
1	Ingest	.754
2	Archival storage	.744
3	Data management	.853
4	Administration	.865
5	Preservation planning	.848
6	Access	.718

As we can see in the above table, Cronbach Alpha for each main functional entity of the model is higher than 0.7. This shows that the checklist stability is favorable.

5. Data analysis

After gathering the existing data at the researcher made checklist, these data were analysed through descriptive statistics (including cumulative frequency percent, average, and standard deviation) and SPSS software. The results are represented at the following tables.

5.1 The most important elements at OAIS reference model Ingest entity

Before Ingest of the data, a data transmission contract (negotiation with works authors about dedication agreements) will be concluded and then the activities of this entity are accomplished, including: obtaining Submission information package (SIP), quality guarantee of SIPs, production of archival information package, extraction of descriptive information from archival information package and coordinate updating.

Among 23 prominent elements at the ingest entity of OAIS reference model, 14 elements with the averages of 5, 4.5 and 4 and with the cumulative frequency percent higher than 80, were identified as the most important prominent elements at the ingest entity, which are presented in Table 2.

Table 2. The prominent elements at OAIS reference model Ingest entity

1. Ingest functional entity		
A transmission contract with dedicator (work author)	Data delivering way (at archives location, through computer network, ...)	
	Determining necessary information at SIP	
	Determining the rate of originality (original or copy)	
Ingest of SIP	Way of SIP delivering	File transfer protocol (FTP)
Quality guarantee of SIP	Controlling of SIP to have no virus before transmission	
	Scanning to control errors	
	Controlling of SIP not to have virus	
Production of archival information package (AIP)		
Extraction of descriptive information from AIP	Cataloging records	
	Index files	
Coordination of updating at archival repository	Making descriptive information integrative	
Data management	transmission of AIP to archival storage	
	transmission of descriptive information to data management	

5.2 The most important elements at OAIS reference model archival storage entity

This entity includes activities such as: the Ingest of archival information package from ingest functional entity and adding it to the permanent memory; hierarchical management of repository; replaced media; determining the rate of the importance of methods for controlling regular and special errors; retrieval improvement; and data acquisition.

Among 21 prominent elements at OAIS reference model archival storage entity, 11 elements with the averages of 5, 4.5 and 4 and with the cumulative frequency percent higher than 80, were identified as the most important prominent elements at archival storage entity, which are presented in Table 3.

Table 3. The most important elements at OAIS reference model archival storage entity

2. Storage functional entity			
	acquiring of AIP from Ingest functional entity and adding it to permanent memory	Determining media type for AIP storage	
	Hierarchical management of repository	Storage methods	Online
		hierarchical	offline
	replace media	Selection of replace media type	
		Transmission methods	Refreshment (like transmission from old DVD onto new DVD)
			Replication (like transmission from DVD onto CD)
	Repackaging (like transmission from GIF format into JPEG)		
	Determining the importance rate of methods for controlling regular and special errors	Allocation of identifiers (ID)	
		Checksum of each data file	
	Retrieval improvement (backing up)	Copying digital content	
	acquisition for information representing	Acquisition a copy from AIP for access entity (for order completion)	

5.3 The most important elements at OAIS reference model data management entity

This entity includes activities such as: management of archival information base, question representation (query), acquisition of report from answers and updating of database (placing of new descriptive information or archival management data).

Among 21 prominent elements at OAIS reference model data management entity, 16 elements with the averages of 5, 4.5 and 4 and with the cumulative frequency percent higher than 80, were identified as the most important prominent elements at data management entity, which are presented in Table 4.

5.4 The most important elements at OAIS reference model Administration Entity

This entity provides services and operations for archival systems and manages the general operations of archival systems. Management tasks include: arrangement of data transmission contract (negotiation with producers about representation agreement), management of hardware and software systems configuration, physical access control, creation of archival standards and policies, activation of cumulative requests, services for users.

Among 28 prominent elements at OAIS reference model data management entity, 23 elements with the averages of 5 and 4; and with the cumulative frequency percent higher than 80, were identified as the most important prominent elements at administration entity, which are presented in Table 5.

5.5 The most important elements at OAIS reference model preservation planning entity

This entity supplies services and operations for monitoring Open Archival Information System (OAIS) and also is responsible for theory provision to guarantee the long-term access to the stored information at

Table 4. The most important elements at OAIS reference model data management entity

3. Data management functional entity		
Archival database management	Creation of descriptive metadata (like Dublin Core and VRA Core)	
	Creation of constructive metadata	
	Creation of preservation metadata (like PREMIS)	
	Creation of legal metadata (like ODRL)	
	Creation of technical metadata (like MIX)	
	Metadata Encoding and Transmission Standard (METS)	
	Studying of information originality and controlling of information integrity	
	Exerting of users' ideas about repository content	
Query implementation	Production of responses set	
	Acquisition of descriptive information	
Acquisition of report set from among responses set	Applying of statistic indicator for access to archival holding	
Updating of database	Updating of descriptive information for new AIP	
	The necessity of system's information updating	Applied statistics
	Checking updating	Checking important information (names and signs)

Table 5. The most important prominent elements at OAIS reference model Administration Entity

4. Administration functional entity	
Arrangement of data transmission contract	Determining of data transmission format at transmission contract
Management of hardware and software systems configuration	The necessity of following up system operations for warranting archival operations
	The necessity of following up system implementation for warranting archival operations
	The necessity of following up system application for warranting archival operations
Updating of archival information	Creation of a mechanism for updating archival content
	Updating through sending dissemination request to access entity
	Updating of the resulted DIP and redelivering it to ingest entity as SIP
	acquiring changes request, procedures and configuration management tools
Exerting physical access control	
Creation of standards and policies	The necessity of determining physical access control policy
Archives	The necessity of determining copy right policy
	The necessity of determining budgeting policy
	The necessity of determining policy for retrieval improvement

	The necessity of determining resource usefulness
	The necessity of determining standards for dissemination formats (according to users' priority, e.g., TIF, GIF, JPEG)
	The necessity of determining pricing policy
	The necessity of determining new archival data formats
	The necessity of determining transmission goals
	The necessity of determining transmission formats
Activation of accumulative requests	The necessity of periodical review of conformity rate of user's request with archives content (in order to assess information availability)
Services to users	Responding to the requested information
	Responding to the users' feedbacks against access services and products

OAIS, even if the environment and the main calculation system is useless. The tasks of preservation planning include: monitoring community of the goal, surveillance technology, developing standards and preservation strategies, developing packaging plans and transmission programs.

Among 25 prominent elements at OAIS reference model data management entity, 14 elements with the averages of 5, 5, 4 and 4; and with the cumulative frequency percent higher than 80, were identified as the most important prominent elements at preservation planning entity, which are presented in Table 6.

5.6 The most important elements at OAIS reference model access entity

The tasks of this entity include: coordination of access activities(methods of user interaction with archives, request type, making resources available, and users validation for access to resources), production of dissemination information package (DIP), and delivering answer to user.

Among 17 prominent elements at OAIS reference model data management entity, 11 elements with the averages of 5, 5, 4 and 4; and with the cumulative frequency percent higher than 80, were identified as the most important prominent elements at access entity, which are presented in Table 7.

6. Conclusion

About 80000 pictures have been so far digitized at the National Archives of Iran. The National Archives of Iran needs strategies for digital preservation in order to supply future generations with access to this digital heritage. "Since reference model is responsible for long-term access to digital resources and has been used for long-term preservation of digital resources by many national libraries and archives, a checklist was designed based on 6 main entities of this model at the present research—including 135 closed questions based on Likert scale (5 optional)—and were given to experts in order to determine the importance rate of each of elements and finally 98 important elements were identified among which 14 elements are related to ingest entity (Table 2), 11 elements to storage entity (Table 3), 16 elements to data management entity (Table 4), 23 elements to administration entity (Table 5), 14 elements to preservation planning entity (Table 6), and 11 elements are related to access entity (Table 7).

Table 6. The most important elements at OAIS reference model preservation planning entity

5. Preservation planning	
monitoring technology	Determining of new digital technologies
	Determining of software detects for new storage and retrieval formats
	The necessity of exerting changes at conditions of services to users (regarding changes at users' information needs)
	Determining of hardware detects for new storage and retrieval formats
Developing of standards and preservation strategies	Migrate
	Transfer
	Emulate
	Technology preservation
	Creation of new standards for dealing with new transmission conditions
	Prediction of changes at conditions of services to users
Developing of packaging plans and migrate programs	Developing of new EEP patterns for responding to migrate goals
	obtaining standards from administration entity and acquisition of SIPs and EIPs
	obtaining new standards for dealing with new transmission conditions
	Applying pre-sample software for responding to migrate goals

Table 7. The most important elements at OAIS reference model access entity

6. Access functional entity		
Through computer network	Methods of user interaction with archives	Coordination of access activities
Interaction through web		
Determining of request type		
(queries, reports, orders)		
Making resources available		
Determining of request posture		
The necessity of user's name	Users validation for access to resources	
User's email address		
retrieving holding (through searching meetings, identifiers on AIP, representing of queries, and returning responses set)		
Determining of request type (set of descriptive information, dissemination information package (DIP), available information resources at archives, reports results)		
Production of DIP (dissemination information package)		
Way of delivering answer to user by archives		delivering answer to user

References

- Anderson, S. et al. "Digital Images Archiving Study." 2006. Retrieved August 9, 2012, from <http://www.ahds.ac.uk/about/projects/archiving-studies/index.htm>.
- Beedham, H. et al. "Assessment of UKDA and TNA Compliance with OAIS and METS Standards." 2006. Retrieved August 9, 2012, from <http://www.jisc.ac.uk/media/documents/programmes/preservation/oaismets.pdf>.
- Nordland, L. P. *The Long and short of IT: The international development research center as a case study for a long-term digital preservation system*. Canada: The University of Manitoba, 2007.
- Samiee, M. "A survey of the status of the digital preservation of the IIPC members' resources and proposing a feasible paradigm to the NLAI." Ph.D. diss., Library & Information Science, Azad University, Tehran, Iran, 2010.
- Samiee, M. et al. (2011). "Survey and analysis of the OAIS reference model." In *Management of digital records*, comp. Gholamreza Azizy, 165-172. Tehran: The NLAI publishing, 2011.
- Semple, N. "Developing a Digital Preservation Strategy at Edinburgh University library." *VINE* 34, no. 1 (2004): 33-37. Retrieved August 9, 2012, from <http://www.emeraldinsight.com/journals.htm?issn=03055728&volume=34&issue=1&articleid=862529&show=abstract>.

Collaboration in Digital Preservation or Lack Thereof: What Works

Digital Preservation in Europe

Strategic Plans, Research Outputs and Future Implementation. The Weak Role of the Archival Institutions

Maria Guercio

Abstract

The European investments for digital preservation in the last decade have been large and persistent but not able to support, at the moment, an accepted common vision, general services and adequate infrastructures. The author investigates this weakness by analysing the nature of the European funding programs in the field and the weak role played by memory institutions, specifically those with archival competences and knowledge, in the European research activities. A specific attention will be dedicated to the relevance for successful results of concepts such authenticity and the role played by a consistent terminology.

Author

Maria Guercio, a full professor in archival science and electronic records management at the University of Rome Sapienza (coordinator for research projects at Digilab Centre), cooperates with the State authority for ICT to define Italian legislation for ERMS and manages ERM training programs for the government school for public administration. She is a partner in many international projects for digital preservation (ERPANET, DELOS, CASPAR), was director of InterPARES' TEAM Italy (1999-2012), and is co-leading the investigation on digital authenticity for the European project APARSEN (2011-2014). She is part of the steering committee of the Section for Archival Education and the Programme Committee of the International Council on Archives. She is author of many articles and manuals in the field. In 2009, she received the Emmet Leahy Award for information and records management.

1. A Premise

The subject here discussed presents a high level of complexity due to the dynamic nature of the technological innovation, the increasing role of the organizational aspects and the interdisciplinary character of the research projects dedicated to digital preservation. To face these difficulties an accurate in-depth and contextualized approach is required. Even if this effort cannot be analysed in detail on this occasion, this conference and its setting offer an extraordinary opportunity for discussing the main promising but also controversial questions in this field. In Vancouver, thanks to UBC's continuing research work on digital preservation, an incredibly productive international environment has been developed in the last twenty years with stimulating results, such as a common terminology, robust conceptual frameworks and significant occasions for international and cross-domain comparisons and advanced educational programs. The UBC's projects—this is an important aspect of their success strictly related to their special original nature—had a very clear archival focus but had and have also the capability of involving other communities with an interdisciplinary approach. Moreover, the questions here proposed for debate have to face new scenarios and new organizational environments which are still undefined: large investments have already been spent or, at least, planned. The main innovative and challenging question is related to the fact that the global financial crisis not only implies new economic difficulties for the research environment but also makes evident the major responsibilities of universities and scientific center, recently defined as *entrepreneurs of innovation* able to “balance the need to articulate a broad strategic vision with the need to execute the day-to-day activities that translate the

vision into reality.”¹ Therefore, the topics here proposed for investigation cannot be ignored, not only for professional and technical reasons, but also and mainly with reference to the centrality in the global society of “research labs, classrooms, and innovations centers, where big ideas are hatched and subsequently translated into reality.”²

Similar to UBC research, many European research initiatives have been largely funded in the same field of digitization and digital preservation in the last decade, but with minor success. They have not been able to maintain a similar level of continuity and, in particular, a comparable level of such global influence. The role and the relevance of the European effort in this research area are undeniable, but they have not been able to re-create (with the partial exception of ERPANET)³ the InterPARES atmosphere of international cooperation and a similar original and authoritative contribution to the research in one scientific domain (in case of InterPARES, the preservation of authentic digital records) and to the enlargement of the interdisciplinary cooperation boundaries. Are these exceptional outcomes due only to the unquestionable quality the researchers involved in the project possess or is something not convincing in the direction and strategies followed in the European funding system in this specific sector? Other successful projects (which are not strictly related to the digital preservation, but connected to the digital heritage) can be mentioned because of their similar global influential capacity as InterPARES: Dublin Core Initiative, the OAIS model, Encoded Archival Description standard, digital curation paradigm, audit framework for trusted digital repositories are just some of the best known products, developed in one sector or thanks to the strict coordination of archival and librarian scholars and professionals and transformed into general and basic elements and tools for building enhanced infrastructures. None of them has been implemented thanks to European funds. Besides, many or all of them have been developed on a domain basis, with a strong commitment of one or two scientific and professional communities. If this did not happen by chance, is it possible to identify and replicate successful conditions as a framework for developing an efficient research environment or, at least, to avoid negative consequences if the requirements are not respected?

These questions are simple and demanding at the same time. Specifically, they concern the identification and the analysis of the main obstacles weakening the relevant investments of the European Commission in the area of digitization and digital preservation and the role played by a cross-domain strategy as factor influencing and partially determining the fragmentation of intermediate and final results of the financed projects. Some assumptions are of course required as a basis for a sustainable line of reasoning.

A first provisional assumption is that the level of success in this complex and dynamic research area was and is linked to the capacity of investigating the domain specific questions with an open approach, based on general principles, solid and consistent concepts and rigorous methods. This means that a common methodological ground is necessary and has to be accepted as fundamental by the research community. The existing scientific traditions, of course, have to be assessed and updated. Well designed and contextualized dictionaries have to be discussed and approved. For this specific reason, a long and cumbersome series of activities are necessary before an investigation program could be operative and fruitful.

¹ Holden Thorp and Buck Goldstein, *Engines of Innovation. The Entrepreneurial University in the Twenty-First Century* (Chapel Hill: The University of North Carolina Press, 2010), 8.

² Ibid., p. 2.

³ “ERPANET - Electronic Resource Preservation And Access Network,” www.erpanet.org (Accessed August 22, 2012).

A second, still provisional, assumption concerns the interdisciplinary nature of the research. Once more, a solid interdisciplinary approach requires highly qualified specialization. The integration of competences can be fruitfully supported only as a second step after the recognition and the comparison of the research values inside each scientific domain.

The theoretical implications behind these two points are very complex and cannot be here investigated in detail. In this specific context, the attention is *simply* dedicated to illustrate the negative consequences (in terms of lack of generally recognized results because these assumptions have been ignored) of the European ‘political’ strategy dedicated to support digitization of cultural heritage and its preservation by facing the technological innovation. This strategy was initially defined in the period 2005-2007 and only partially updated in 2011⁴ with the goal of financing mainly (if not solely) interdisciplinary projects explicitly focusing on digital convergence. But the nature and the content of this interdisciplinary approach and the meaning and consequences of the digital convergence have never been defined and the convergent inclusion has excluded many relevant scientific communities and domains. The archival sector, in particular, even if its digital heritage is one of the most challenged by present and future technological innovations, has been substantially ignored by the European investments. One of the reasons is related to the underestimation of the potential negative potential impact and consequences of this marginalization (not only for the domain itself) in terms of advanced solutions for the research and for services implementation.

As a matter of fact, the limited contributions made available to the research activities of the European archival community (as will be further illustrated) have prevented its institutions from participating efficiently in the investigation regarding the main innovative topics. Therefore, the great European documentary traditions, diversified but also consistent in their basic common principles and methods, could not have been recognized and exploited in Europe and internationally with at least two relevant limits for the global (not only European) research. Firstly, it has been impossible to develop at the European level a strategy toward a systematic and standardized approach for electronic recordkeeping and preservation systems.⁵ Secondly, this delay has prevented the archival and record management community from providing effectively (that is as a recognized and authoritative network) its contribution to the research in the specific field of digital preservation and to the definition of digitization standards and parameters. The low level of qualified convergence and satisfactory level of interoperability of digital library initiatives (including Europeana, the EU digital library) provides undeniable evidence for these limits.

Since its first steps, and especially after 2007, the European strategies for the digital heritage have marked the centrality of actions aimed at promoting the cultural contents convergence, by overcoming quite always the specificities in the research investments and favoring (and selecting) quite only projects able to guarantee the largest participation of cultural institutions of the widest and most variegated nature

⁴ The relevant European reports on these aspects are: “i 2010 Digital Libraries Initiative” (2005), Commission Recommendation on Digitisation and Online Accessibility of Cultural Material and Digital Preservation (2006 – updated October 27, 2011), “Member States’ Expert Group on Digitisation and Digital Preservation (2007), “Communication from the Commission: ‘Scientific Information in the Digital Age: Access, Dissemination and Preservation’ (2007), Recommendation on Digitisation and Digital Preservation (October 2011).

⁵ The case of European recommendations for electronic record management systems, known as MoReq (MoReq1, MoReq 2 and MoReq2010), is emblematic: the specifications have required 10 years to be approved and to develop certification process but the final version of MoReq2010 is still under development, services for certification are not in place: and the European organizations at the moment are not compliant with them. See M. Guercio, “MoReq1, MoReq2 e MoReq2010: raccomandazioni e prove tecniche di certificazione per la gestione informatica dei documenti,” *Archivi* 1 (2012): 7-32.

and provenance as possible: as mentioned, the focus of all the European recommendations for digitization and digital preservation was and is on convergence of domains for accessibility and for interoperability. The quantity but not the quality is crucial for this strategy. At a matter of fact, the explicit political goal in 2007 and, even more, in 2011 was and is to ensure a critical mass of contents for the European digital library, called Europeana by:

[...] setting up or reinforcing national aggregators bringing content from different domains into Europeana, and contributing to cross-border aggregators in specific domains or for specific topics, which may bring about economies of scale, ensuring the use of common digitisation standards defined by Europeana in collaboration with the cultural institutions in order to achieve interoperability of the digitised material at European level, as well as the systematic use of permanent identifiers.⁶

The challenges involved in these new services risk to underestimate the role of the cultural heritage scholarships and knowledge differentiation and specialization which were and are at the basis of the European quality. “Convergence against integration” can appear as an easy and rapid shortcut for dealing with the digital convergence processes, but not necessarily will provide efficient solutions. In our specific area, this general policy had the consequence, among other things, of pushing the research institutions (and of course also the institutions of memory) to present projects characterized by a high degree of heterogeneity of purposes, participants and research components and activities. Even if nobody in any domain explicitly contested and contrasted this perspective and this orientation, the final result has been the marginalization of all the sectors not really involved in digital access and dissemination of published and native open public materials (that is digital resources immediately available on the web), such in the case of e-health records, or in the private business environment and in general in the recordkeeping systems).

Of course, as a consequence, any possibility for scientific investigation in each domain has been sacrificed to ensure multiplicity and diversity of actors, of contents and vocabularies and, more recently, to increase the quantity of information available and not the quality of its representation in terms of contextualized relations (that is with reference to the digital objects intelligibility and the capacity to assess over time their integrity and authenticity, to trust both digital contents and their relations). Even more, the illusion that the complexities involved in the digital preservation research could find at the end exhaustive answers in the multidisciplinary comparison has influenced the whole research sector: the partnerships could not be developed on the basis of the specific qualifications of the institutions to involve. The domain diversity of the projects partnerships was and still is an obliged requirement for successful funding. It has to be said that the institutions of memory (mainly the archival organizations) have shown here their limited capacity for developing strategic alliances (relevant not only for funding and for exploiting finance channels but mainly to create strong and permanent institutional interconnections. In the specific archival sector, another negative factor has been the substitutive role of coordination assigned, on an unclear basis and with insufficient commitment, to organizations like DLM Forum Foundation.⁷ This association, born in 1994 as “an inter-disciplinary cooperative effort led by the

⁶ Commission recommendation of 27.10.2011 on the digitisation and online accessibility of cultural material and digital preservation, Accessed August 21, 2012, http://ec.europa.eu/information_society/activities/digital_libraries/doc/recommendation/recom28nov_all_versions/en.pdf.

⁷ The recommendations of the European Council of November 14, 2005 were limited to ask the initiative of the Commission on the following aspects: “further development of European interdisciplinary co-operation on

EU member states and the European Commission,” became in 2002 an independent body and, in 2010, a non-profit foundation in the area of “archive, records, document and information lifecycle management throughout Europe.” More and more the Foundation has seen the increasing interests and direct participation of software companies and has diminished its original potential nature of a neutral network based on institutional values. Even if recognized by the European Commission, the organization has never been strongly involved in the European research projects. The role assigned by the Commission to the European Archival Group which (will be analysed more deeply in the second part of this presentation) had a similar marginal nature for the opposite reason: it has been supported for coordinating archival actions among the European institutions but it has never been included in the core research projects dedicated to the technological innovation, specifically preservation and digitization. The only exception is the APENET and now APEX projects and the portal dedicated to make the archival data available on the web and cooperate for ensuring interoperability with Europeana.

It is surprising to observe that the international scenario shows less archival marginality than the European context, even if (or because?) the work is done on the voluntary basis and the resources are very limited in comparison with the large amount of European investments. It includes the work done under the umbrella of the International Council on Archives, or the international research projects such as InterPARES and the projects on trusted digital repositories, but also the working groups of ISO committees and subcommittees on archival and record management (TC 46 SC 11). The archival Europe supported at various levels (mainly with direct participation) all these efforts thanks to the commitment of single institutions, but these actions and their strategic objectives have been substantially ignored by the EU funding bodies, as the available information on the last decade funded projects clearly illustrates:

- With exclusive reference to the research on digital preservation for the period 2006-2011, fifteen projects have been financed under ICT programs (the most important in the field here investigated); they have been funded with 86 million Euros with the specific aims of creating technology solutions and innovative methods to keep *digital resources available and useable over time*,⁸
- In 2012 other 30 projects have been approved for the amount of other 30 million Euros;
- None of the funded projects had and has an archival leadership (academic or institutional);
- The institutions with a specific and recognizable archival competence involved in these projects are very few (5-6), while the European partners involved are hundreds: as far as I know, Institution for Archival Studies at the University of Urbino (Caspar⁹), the National Archives of the Netherlands, the National Archives of England, Wales and the United Kingdom, the Swiss

electronic documents and archives, updating and extension of the requirements for setting up electronic document and archive management systems, such as MoReq and a continuation of the DLM Forum conferences on electronic documents and archives”. For the history of FLM Forum Foundation see http://www.dlmforum.eu/index.php?option=com_content&view=article&id=13&Itemid=15&lang=en. Accessed August 22, 2012.

⁸ Pat Manson, *Stakeholders’ Workshop on eArchiving* (Luxembourg 24 February 2012).

⁹ “Caspar - Cultural, artistic and scientific knowledge for preservation, access and retrieval”, 2002-2006, Accessed August 20, 2012, <http://www.casparpreserves.eu>.

Federal Archives (Planets),¹⁰ the National Archives of Sweden and the National Archives of Estonia (Protage).¹¹

In contrast with the recent evolution, the first European strategy for digital preservation was, at least apparently, differently oriented. The ERPANET project, a *network of excellence* active in the period 1999-2003, was run by only 4 partners. Three of them had archival competences (the National Archives of Netherlands, the National Archives of Suisse, and the Institution for Archival Studies at the University of Urbino). The principal coordinator, Seamus Ross, director of HATII at the University of Glasgow, was at that time a close observer of progress made by the documentary disciplines in the field. The same partners (only the National Archives of Suisse were not included) successfully collaborated in the DELOS¹² project within the same work package dedicated to the digital preservation and developed promising ideas which were further implemented by following projects such as PLANETS¹³ and DPE¹⁴ (among other results, a set of procedures to document self-auditing with reference to the requirements for digital repositories trustworthiness, the systematic analysis of file formats for preservation, the creation of test-bed environments).

In the recent projects the archival dimension of the research progressively became marginal and sporadic. Other domains have been able to develop their influence and assume a strategic role in cooperation initiatives such as the library sector with reference to the European digital library Europeana¹⁵ or the network of government cultural institutions which still play a coordinating role thanks to a continuing series of EU funded projects (Minerva and Minerva eC,¹⁶ Michael¹⁷ and, more recently, Athena¹⁸ and Linked Heritage¹⁹). Also the audiovisual institutions were strongly supported in Europe with dedicated programs able to create a healthy chain of continuity (Presto, Prestospace²⁰ and Prestoprime).²¹

How effective is the cooperation in these cases is a question assessed only when the single project is evaluated by the Commission in the course of the funding process. At the moment there are no mechanisms in place (but only political initiatives) to develop on more general basis assessments of research areas and/or financing lines, or to evaluate and solve critical questions like the lack of interactions and cooperation or integration among sectors, even if explicitly recognized by official documents.

¹⁰ "Planets - Preservation and long-term access through networked services", 2002-2006, Accessed August 20, 2012, www.planets-project.eu.

¹¹ "Protage", 2010-2012, Accessed August 20, 2012, www.protage.eu/project.html.

¹² "Delos Network of Excellence on Digital Libraries", 2004-2008, Accessed August 20, 2012, www.delos.info/.

¹³ Planets - Preservation and long-term access through networked services, Accessed August 22, 2012, www.casparpreserves.eu e www.planets-project.eu.

¹⁴ "DPE - DigitalPreservationEurope", Accessed August 22, 2012, www.digitalpreservationeurope.eu.

¹⁵ "Europeana", Accessed August 20, 2012, www.europeana.eu/portal.

¹⁶ "Minerva" and "Minerva eC", Accessed August 20, 2012, <http://www.minervaeurope.org/>.

¹⁷ "Michael", Accessed August 20, 2012, <http://www.michael-culture.org/it/home>.

¹⁸ "Athena - Access to Cultural Heritage Network Across Europe", 2008-2011, Accessed August 20, 2012, <http://www.michael-culture.org/it/home>.

¹⁹ "Linked Heritage", 2011-2013, Accessed August 20, 2012, <http://www.linkedheritage.eu/>.

²⁰ "Presto", 1999-2002, Accessed August 20, 2012, <http://presto.joanneum.ac.at>; "PrestoSpace", 2004-2007, Accessed August 20, 2012, www.prestospace.org.

²¹ "Prestoprime", Accessed August 20, 2012, www.prestoprime.org

It is the case of the official critical remarks (published in the final report of the project) about the low level of inter-sector cooperation verified and recorded by the archivists involved in the APANET project when a *convergence* process was activated to create compliance and interoperability between the descriptive systems and standards of archival sector and those already developed by Europeana:

This is the final report of the APEnet WP3 team in which it describes the interoperability efforts it undertook throughout the whole APEnet project period as well as the difficulties it came across and the ways it found to deal with them [...]. The overall conclusion of this report is that the APEnet WP3 team has succeeded in making the technical interoperability fully operational, so in fact has done the job it was supposed to do, but could have accomplished more in case the Europeana team would have been more cooperative. The reason for this is that – despite a lot of efforts from the APEnet WP3 team – the interoperability on a strategic level has been a one way communication most of the time during the entire APEnet project period, plus the fact that on a technical level Europeana was much like a moving target and at the end of the APEnet project even out of reach for the APEnet WP3 team.²²

The fact is that the European archival institutions, whose historical custodial function could have supported, with the knowledge and experience accumulated for centuries, the understanding and the translation in the technological environment of crucial concepts such as digital trust, reliability, accuracy and authenticity and identify implementation methods and tools in the field of digital preservation, have never acted as main characters or leading protagonists. They obtained merely *protection* but not recognition as crucial players for central challenges. They have not yet created an effective European network able to support research and initiatives related to digitization processes and digital preservation. The DLM Forum Foundation²³—as has been already observed—lacks the institutional status for playing this role. The only initiative, lightly supported by the European Commission and previously mentioned, is the *European Archives Group*²⁴ but its attention is mainly dedicated to the traditional and still complex challenges of the analogue archives.

In the specific area of European research for digital preservation, a general and relevant critical question concerns the brief funding period always granted to the projects, no matter how promising the initiatives are: according to this policy, each research project should be able to survive and become sustainable in only 3-4 years of financial support and should be able to develop its own research center for strategic future programs. In other scientific sectors (as in the so-called *basic research*) European support ensures more continuity of funding and, for this reason, makes success and international recognition more possible.

The fragmentary nature of the European support for digital preservation research—partially mitigated by the not unusual opportunity for an institute and its researchers of being funded for years even if in different projects (as happened in the case of University of Urbino)—has strong and specific political

²² APENET, “Final Interoperability Report Deliverable D.3.2”, Accessed August 21, 2012, www.apenet.eu.

²³ “DLM Forum”, Accessed August 20, 2012, <http://www.dlmforum.eu>.

²⁴ “European Archives Group – EAG”, Accessed August 18, 2012, http://ec.europa.eu/transparency/archival_policy/eur_arch_group/index_en.htm. The working group is based on the participation of National archives directors. The main activities have concerned the creation of a European archival portal (Apenet, now Apex), the database of European archival legislation Euronomos (in cooperation with the ICA European Branch EURBICA), the definition of guidelines on the thefts of archival materials and other studies. Relevant goals and actions, but not sufficient for building a strategy and a solid and persistent infrastructure.

EU reasons, which cannot be easily overcome even if in the European Research Framework new perspectives have been identified and could be able to create a radical and positive evolution in the near future. It is moreover clear that many other factors contributed to maintain the present situation, specifically characterized by isolated projects and by difficulties in creating fruitful interconnections and supporting an effective strategy:

- The market and the stakeholder interests are relevant, variegated and conflicting and, also for this reason, the differentiation of the investments is prevailing in the evaluation process;
- Archival institutions, which are involved (even if they are not numerous) in the European projects, have not been able to develop a common strategic leadership and promote a coordinating action; the other institutions which have not been involved not seem even aware of their exclusion in the last decade of European funding.

With particular reference to the last point and to the obstacles which have further limited the effective cooperation in Europe among archival institutions, many elements have to be considered. They include:

- The institutional resistance and reserve about working on single projects without insurance for continuity, while (or because) they are facing other dramatic, still open and unsolved challenges (such as the documentary proliferation and the complex problems related to e-government initiatives both in the legislation and in the national governmental implementation);
- The difficulty of and the delay in conjugating the institutional activity for protection and custody of archival sources with the scientific investigation and in cooperating with the research initiatives conducted by the academic institutions;
- A traditional autonomous attitude attested in the past by the delay in supporting standardization processes;
- A passive acceptance of the European policies established in the period 2005-2007 focused on principles of convergence and interoperability of the digital cultural heritage;
- The underestimation at European and at national level of the problems related to the preservation of digital archives and to the need of building an early and consistent network of custodial institutions, which are authoritative, reliable, certified, trustworthy and neutral. It should be sufficient to remember the very recent proposal of the European Commission to modify the regulation related to the European historical archives²⁵ with reference to the removal of the obligation of transfer to the central archival repository in Florence for the European institutions when the records are digital: in this case the creators, according to this new proposed rule²⁶ and against any consolidated archival principle, will maintain their complete responsibility for ever (“the originating institutions will remain responsible for the long-term preservation of their digital

²⁵ “For a Council regulation amending, Reg. (EEC/Euratom) No 354/83, as regards the deposit of the historical archives of the institutions at the European University Institute in Florence”, Accessed at August 24, 2012, <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:DKEY=687809:EN:NOT>.

²⁶ The annex related to the provision for the deposit of the historical archives of the European Union institutions makes only explicit that “in the case of digital archives, the EUI shall have permanent access to the documents in such a way as to allow it to fulfill its obligation to make the historical archives accessible to the public from a single location and to promote their consultation”, *ibid*.

archives”). Because no other rule is defined with respect to this responsibility, the Commission proposal seems to completely ignore the risks of losing control over the quality and the integrity of the records when they are no longer active (and relevant for the creators) and are not yet (and perhaps will never be) transferred to dedicated *trusted* archival repositories. As stressed by Angelika Menne Haritz, when the proposal was discussed at the EAG meeting in October 2011, because “digital archives are not fixed,” only the custody in archival repository assigned to third neutral parties is able to guarantee that “these archives will not change after they are opened to the public.”²⁷ The decision of the European Commission is in patent contradiction with the main results of the research projects funded by the Commission itself, explicitly in favor of the creation of certified institutions dedicated to the digital preservation as the essential requirement for long-term preservation.

From this point of view the archival institutions, the absence of archival competences and knowledge in the European research for digital preservation determined, among other consequences, that many research questions, which are still relevant and critical for the domain, have not yet found an adequate and systematic analysis and risk remaining unanswered (and sometimes also unexpressed). Only some of them can be here listed:

- How should one apply OAIS in the archival domain?
- How should one transform the archival requirements for information representation and the specific references to contexts and complex relations into the digital library standards without losing the intelligibility of the digital objects made available on the web?
- How should one develop services able to document the custodial history in the transfer processes?
- How and when should one document the authenticity evidence and who would be responsible for its maintenance?

These questions (and many others) are urgent and cannot find adequate answers if the institutions are left alone. It is not by chance that recently, thanks to the request of some European governments and, specifically, of Slovenia, a specific action concerning *long-term data archiving* has been finally proposed to the European Commission (*Proposal for support of eArchiving in CIP ICTPSP WP12*) and partially accepted. The focus is—at the moment—limited to the transfer of custody for digital archives disposed for long-term preservation as proposed by Jože Škofljanec (Archives of the Republic of Slovenia) at the first meeting. This meeting was held in Luxembourg in February 2012 with the participation of many archival stakeholders and experts for digital preservation and the proposal was presented with the following motivations: “the level of digital preservation is very different among member states; (limited) practical experience is available; pan-European standardisation is taking place in the field of RM (MoReq) but not regarding ingest procedures.” The Slovenia proposal was promptly supported by all the archival institutions of many European countries (among others Austria, Denmark, Estonia, Finland, Italy, Hungary, Latvia, Netherlands, Norway, Portugal, Spain, Sweden, UK) and the experts involved in this internal workshop. The project obtained the recognition of its relevance when the government representatives met some months later. The basic request concerned the increase of efforts and support for

²⁷ European Archives Group – EAG, Minutes, accessed August 18, 2012, http://ec.europa.eu/transparency/archival_policy/docs/eag/cr/111007_minutes_en.pdf.

developing guidelines and common tools (that is, fewer projects and more coordinating actions, programs and working groups). Specifically, the final report made explicit the need for “common set of guidelines and tools for transfer and ingest processes and formats in line with European interoperability initiatives” to make available to the member States “simple pan-European access tools and e-services – strengthen reuse.” A pilot project is now planned in the form of a call for a proposal for “developing and piloting eArchiving service solutions, the necessary technical infrastructure, policies and processes to support the preparation and transfer of content and its ingest to repositories, ensuring its authenticity, provenance and integrity over time as well as its continued availability and usability”—in synthesis for implementing “scalable sustainable practical eArchiving services.”²⁸ Of course, the recognition of the critical value of this topic does not imply a correct archival solution. The risk of delegating the main preservation problems (such as authenticity and integrity) to technological mechanisms is always around the corner, as was made evident by the Italian legislation approved in 2004.²⁹

2. The methodological contribution of the archival knowledge for digital preservation research

From the methodological point of view, the most relevant open question concerns the concrete and specific content of the contribution that the archival domain should be able to provide to the research dedicated to the digital preservation. The answer can be found by analysing the results of the European researches and, specifically, their main weaknesses. Of course, there are no parameters for identifying such deficiencies and no metrics available to support such analysis. Evidence for this risk can be found in the APARSEN deliverable for work package on authenticity and interoperability.³⁰ On this occasion the researchers had to revise dozens of deliverables and reports published by recent European projects dedicated to digital preservation and were able to verify how limited at the moment the investments (internal to the projects) are for the definition of a project’s stable vocabulary and for a systematic conceptual framework³¹ and how difficult is to orient users without such tools which be able to function as *protocols of understanding*.

On the contrary, both the terminology and the conceptual framework are fixed ideas, even an obsession, for the archival community, specifically when they operate at the international level. There is no doubt about the capacity of the documentary disciplines and, in particular, the archival science, to provide effective contribution to this activity, specifically for the creation of a theoretical outline. The international experience matured by the archival community on this subject is long lasting and well qualified (as it will be later explained).

²⁸ Cfr “Orientations for CIP ICT PSP WP2013” (internal document) and in particular the point dedicated to “Scalable sustainable practical eArchiving services”.

²⁹ See Maria Guercio, *Archivistica informatica* (Roma: Carocci, 2010), chapter 5.

³⁰ APARSEN Project: Deliverable 24.1. Report on Authenticity and Plan for Interoperable Authenticity Evaluation System, 2012, Accessed August 22, 2012, http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D24_1-01-2_3.pdf.

³¹ It is frequent to meet new and specific glossaries, created for each deliverable and developed by simple accumulation of terms of various provenances. Even the compliance with recognized standards is not always required or pursued. Of course, building and sharing vocabularies is a very complex and ambitious task: it is never a question of collecting terms, but it implies a strict control of the relations and, more important, their contextualization. The agreement among co-authors is also a very demanding work.

With reference to the conceptual framework, someone could object that digital preservation systems already follow a standardized reference model, which is at the same time both consolidated and flexible, that is the OAIS model. This model is very important and generally appreciated both when general rules have to be developed (see the case of Italian legislation under approval³²) and when a digital repository has to be created and implemented (as in the case of Regione Toscana which is presented at the conference by Ilaria Pescini paper) or reconsidered (again, as in the case of e-health standard for Italian government, UniSInCRO³³). This model is today and will remain for a long-time the benchmark for digital preservation projects, but (also because of its nature necessarily open, flexible and abstract), it is unable to meet more specific requirements. The critical aspects regard not only the sectorial, domain-dependent vocabularies, but also the conceptual structure needed for sustaining in the daily life of the digital resources the implementation of the main categories of Preservation Descriptive Information (provenance, reference, context and fixity). These categories are very often difficult or impossible to handle separately. A robust methodology is required to avoid ambiguities and overlapping.³⁴ It is not by chance if the just published release of OAIS has reinforced its relationships with some archival concepts, such authenticity, even if it developed in a peculiar manner.³⁵

In particular, the authenticity is a very crucial subject, increasingly recognized for its centrality among the concepts commonly used and referred to by the communication, information and knowledge society. The information society by its nature directly works on the creation and narration of social and individual identities and, for this reason, requires tools, procedures and fundamentally robust concepts for entrusting and documenting their authenticity specifically when facing the digital world challenges. Moreover, the authenticity in the sense of identity of the resources and their integrity has not by chance been developed by the documentary disciplines. It has been creatively and thoroughly analysed by Paola Carucci³⁶ with reference to the contemporary records and investigated for the digital environment by Luciana Duranti and other InterPARES researchers.³⁷ The knowledge and the experience of archival

³²“Regole tecniche di conservazione di documenti digitali”, <http://www.digitpa.gov.it/sites/default/files/Bozza%20-%20Regole%20tecniche%20conservazione.pdf>, accessed September 5, 2012.

³³ Uni 11386:2010 – “Supporto all’interoperabilità nella conservazione e nel recupero degli oggetti digitali (SInCRO)”, Accessed August 22, 2012, <http://webstore.uni.com/unistore/public/searchproducts>

³⁴ To make these concepts practically applicable in preservation cross-domains environment, a double exercise is required. First of all it is necessary to translate/accommodate them into the OAIS framework, specifically into the information package relevant for handling digital content preservation, the Preservation Description Information which includes Reference, Context, Provenance and Fixity Information. Of course, this ‘translation’ is not automatically extensible: an interpretation is required based on the definition of resource typologies and linked to specific domains knowledge. Some contradictions among OAIS PDI categories still require to be solved, as in case of the reference information: for archival records this concept implies identifiers of various levels (persistent identifiers, but also registry, classification code) and includes context information like archival bond while OAIS defines context as a separate area of PDI (the relationships of the Content Information to its environment, i.e., why it was created, how it relates to other Content Information objects, etc.).

³⁵ CCSDS: Reference Model for an Archival Information System – OAIS. Draft Recommended Standard, 650.0-P-1.1 (Pink Book), Issue 1.1, 2009, Accessed August 22, 2012, <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/CCSDSAgency.aspx>.

³⁶ Paola Carucci, *Il documento contemporaneo*, Roma: Carocci editore, 1987.

³⁷ See, among other, Luciana Duranti, Terry Eastwood, Heather MacNeil, *Preservation of the Integrity of Electronic Records* (Dordrecht: Kluwer Academic Publishers, 2002), Chapter 1; Luciana Duranti, Heather MacNeil, ‘The Preservation of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project,’ *Archivaria* 42, no. 2 (1997): 46–67; Heather MacNeil, *Trusting Records. Legal, Historical and Diplomatic Perspectives* (Kluwer Academic Publishers, Dordrecht, 2000).

institutions and scholars are in this respect undoubtedly central and the concept of authenticity is undoubtedly central for the research on digital preservation and for building measures and tools for trusted digital repositories of any type.

Another sector where the archival contribution is substantial and unavoidable—as, at the end, was also by the information scientists involved in the research environment—concerns the use of principles, concepts and tools borrowed from the *record management*. This also includes recommendations and standards developed in the course of the last decade by ISO committee on archives and record management and by ICA. Among other things, they concern the recognition that the digital preservation, intended as dynamic process for digital continuity, can be sustainable and can be handled only on the basis of information lifecycle management or in the form of digital continuity because of standardized and well controlled vocabularies consistent with the functional model which governs the information phases and layers. From this point of view OAIS is the essential reference model, but—as previously underlined—is not self-supporting when consistency of contents is required.

With reference to the global research on digital preservation, conceptual frameworks, well defined vocabularies, contextualized concepts should be further developed and discussed if a real and effective interdisciplinary approach would be supported for improving the general investigation methodology. They are fundamental requirements for any successful project, specifically when the sector is undetermined and in continuing transformation as the digital environment is. As was underlined years ago by Seamus Ross in his keynote speech at the Budapest conference on digital libraries, the digital contents in any domain “require knowledge of its context of creation, and they demand evidence of its provenance. These are processes to which archives respond well because they have developed an appropriate theoretical framework and have operationalised it in repository design, management and use over at least three centuries. The archival framework meets requirements surrounding the production, management, selection, dissemination, preservation and curation needs of information. It also supports a layering of services from repository services at the foundation to user services at upper levels.”³⁸ As been noted, this archival framework has been further developed and standardized in the last decade both with reference to the vocabulary and to the functional model by the theory of electronic recordkeeping and the standardization processes developed in this sector, thanks to the cooperation of a large international community based on ISO and ICA initiatives, has provided a significant contribution to support the collection of relevant information for integrity and identity of digital resources in the creation and preservation processes. In general, this effort (which is not necessarily linear, but on the contrary is very often conflicting and contradictory) is still operative and implies working groups at national and international level and thorough knowledge from each domain.

With specific reference to the authenticity, a rich archival literature has been developed, based on the InterPARES research and its *template for analysis*. Nevertheless, not many European projects have used this reference structure. The only exceptions have been CASPAR and now APARSEN³⁹ which have

³⁸ Seamus Ross, “Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries” (keynote speech presented at ECDL Budapest 2007), p. 8. Accessed August 18, 2012, http://www.ecdl2007.org/Keynote_ECDL2007_SROSS.pdf.

³⁹ This analysis has been discussed by the author in the deliverable D24.1 of APARSEN project: Silvio Salza, Mariella Guercio, Monica Grossi, Stefan Pröll, Christos Stroumboulis, Yannis Tzitzikas, Martin Doerr, Giorgos Flouris, “D24.1 Report on Authenticity and Plan for Interoperable Authenticity Evaluation System,” April 2012, accessed August 18, 2012, http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D24_1-01-2_3.pdf.

integrated, thanks to the presence of archivists involved in the research and of course with the support of experts of OAIS functional model, the template elements into OAIS. The other European projects, even those which had the authenticity management as part of their crucial requirements, have not proposed original solutions and have normally ignored the question and its definition. This lack of interest is not easy to understand and even more is difficult to justify, specifically because today terms such *trust* and *reliability*, at the centre of any research and any debate on digital preservation, acquire their intelligibility and resonance, if and when they can be concretely measured against the capacity to verify or, at least, presume the authenticity of the digital resources to be preserved. The same question is at the basis of any digital certification process of digital repository.⁴⁰ The question of authenticity and how the research projects are or not normally able and/or willing to handle it can be emblematic of the present tendency to overcome the critical factors by ignoring them, instead of sustaining the effort to identify the nature of the problems, the complexity of the actions involved and the related responsibilities on the basis of the operational contexts.

The need for consistent and reliable vocabularies is again central (but of course cannot be handled as an accumulation of terms). They make the research results intelligible, measurable and exchangeable and allow for the continuity of the investments and the research work done. As already has been said, the archival community had a strongly commitment in this area. The promotion of a consistent archival terminology has been, not by chance, a key point of the initial program for ICA when launched by the director of US National Archives, Solon J. Buck, in 1946 and still is at the center of ICA attention.⁴¹

3. The future of the European action: New opportunities within the Common Strategy Framework

The European bodies are certainly aware of the limits and of the challenges still open at the present state of their investments and which are here only partially listed and discussed. One of the main qualities of their policies is the capacity to continuously update their high level evaluation processes. More complex is the transformation of the intermediate decision level, specifically when the challenges involved are even difficult to be defined and to be designed for future strategies. An important (even if not decisive)

⁴⁰ See on this aspect APARSEN Project: Deliverable 24.2. Implementation and testing of an authenticity protocol on a specific domain, 2012, Accessed August 22, 2012, http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D24_2-01-2_2.pdf.

⁴¹ The Buck roll call, titled *A Proposed Archives Program for the United Nations Educational, Scientific and Cultural Organization* (September 1946) is published as annex in the article of Oliver W. Holmes, "Towards an International Archives Program and Council, 1945-1950," *American Archivist* 39, no. 3 (1976): 296-299. See on this aspect Giulia Barrera, "Il villaggio globale degli archivisti. Organizzazioni internazionali e forme di cooperazione tra gli archivi del mondo," in *Archivistica*, ed. Linda Giuva and Maria Guercio (Roma: Carocci, in press). The ICA has dedicated time and financial resources to develop and update its dictionaries. The first dictionary was published in 1964: *Elsevier's lexicon of archive terminology: French, English, German, Spanish, Italian, Dutch*, compiled and arranged on a systematic basis by a Committee of the International Council on archives (Amsterdam - London - New York : Elsevier, 1964). A second series of tools has been published in many languages: *Dictionary of Archival Terminology: English and French, with Equivalents in Dutch, German, Italian, Russian, and Spanish = Dictionnaire de terminologie archivistique*, ICA handbooks series, 7, 2nd rev. ed., ed. Peter Walne (München - New York: K.G. Saur, 1988). A third edition completely renewed with contextualized definitions developed in 2004 was only published on the web by Angelika Menne-Haritz, Accessed August 20, 2012, <http://www.staff.uni-marburg.de/~mennehar/datiiii/intro.htm>.

step in this process was the EU workshop held in Luxembourg (Shape New Visions for EU-Research On Digital Preservation, May 4-5, 2011),⁴² with the participation of many European and North American researchers and experts in the field, invited for defining the new vision for EU research on digital preservation. The final report recognized that:

There was not yet a corpus of best practice to refer to. These communities were not, however, well connected, and had not yet engaged widely beyond public memory institutions and research bodies. There also appeared to be lack of dialogue between digital preservation specialists and those who produce and curate data.⁴³

The results of this workshop, by recognizing (even if partially) some of these needs, suggested:

- Shifting the attention from the preservation of data to the preservation of knowledge: this shift would be “be able to demonstrate the authenticity of preserved data,”
- Focusing more on the integration of digital preservation into lifecycle management of digital contents as proposed a decade ago by the record management and archival science scholars.

These two conclusions seem to confirm what was previously observed with reference to the areas where the contributions from the archival community could have been more than productive. The workshop provided many other opportunities for discussing all the open questions in the field, even if many aspects were not debated at all, specifically those of strategic value. For instance, the report in any way did not include, in the list of missing actions, the creation of reference centers or networks for the definition of rules, services, procedures, tools and also methods, even if similar ideas are present in projects like APARSEN (Virtual Center of Competence)⁴⁴ or in PLANETS (Open Planets Foundation).⁴⁵ The eArchiving pilot proposal (see above) can also be interpreted as a first recognition of this weakness, in this case specifically referred to the archival domain. It will be the responsibility of the European archival institutions to transform this tepid credit opening into a real opportunity and an occasion for collecting

⁴² “Shape New Visions for EU-Research On Digital Preservation,” Luxembourg, May 4-5, 2011, Accessed August 20, 2012, http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-future-digital-preservation_en.html.

⁴³ Clive S.G. Billness, “Report on the Proceedings of the Workshop The Future of the Past. Shaping New Visions for EU-Research On Digital Preservation,” organized by Cultural Heritage and Technology Enhanced Learning, Luxembourg, May 4-5, 2012. Accessed August 19, 2012, http://cordis.europa.eu/fp7/ict/telearn-digicult/future-of-the-past-summary_en.pdf. See also the report presented at the workshop, Stephan Strodl, Petar Petrov, Andreas Rauber, “Research on Digital Preservation within Projects co-funded by the European Union in the ICT Programme, 2011,” Accessed August 20, 2012, http://cordis.europa.eu/fp7/ict/telearn-digicult/report-research-digital-preservation_en.pdf. The report presents an overview of the past and present European research projects. It is interesting to note that it is recognized that in the first phase (Delos and ERPANET) a significant effort was developed - under the influence of the library and archival communities - for the “establishment of common problems, definitions, terminology and concepts” (p. 52), while in a second phase the attention was (and is) concentrated on non-traditional objects and on more technical aspects, with less attention to the development of common bases. By illustrating the outputs and topics of the funded European projects, the report makes evident, even if not explicitly, the fragmentation of the whole picture. It is useful to stress that only two archival experts were invited to the workshop and the awareness that this absence could have produced a negative impact on the definition of a more complete strategy was not at all expressed in that occasion.

⁴⁴ “Aparsen - Alliance for permanent access to the records of science network,” Accessed August 22, 2012, www.alliancepermanentaccess.org/current-projects/aparsen.

⁴⁵ “Open Planets Foundation,” Accessed August 22, 2012, www.openplanetsfoundation.org.

ideas and proposals and creating the basis of an effective network or of a virtual center of archival competence.

Another important initiative, which should attract the future attention of the stakeholders, is the new European financial infrastructure proposed for the period 2014-2020, specifically with reference to the definition of a Common Strategy Framework for innovation and research and to the future plans included in the program called Horizon 2020.⁴⁶ In relation to the aim of “Connecting Europe Facility (CEF)—cross-border infrastructures in energy, transport and ICT to strengthen the internal market” the new program defines the need to reinforce with adequate research resources “*competence centres on digitisation and preservation of digital cultural heritage*.” The European and international networks for cooperation are here indicated as strategic tools. These changes could provide new possibilities for overcoming the old critical difficulties but the challenges (which were not faced in the past as has already been indicated) will be even more complex in the near future and will require great capacity for cooperation, clear ideas and qualified alliances. The present is characterized both by the instability of the technological innovation and the complexity of the creation of digital contents, but also by the continuing growth in the development of digital data which is “outstripping the rate of growth in data storage technologies.” This uncertainty can be faced or at least limited with contextualized definitions of solid basic concepts and with a persistent coordination among qualified institutions, possibly characterized by common goals, even if the dynamic nature of the digital environment will never stop in provoking new conflicts and creating new uncertainties. The digital environment imposes (not only in case of digital heritage) undefined boundaries and *seems* to make irrelevant the traditional specificities, but implementations in the real world does not accept a high percentage of ambiguity as has been increasingly recognized in the recent debate on postmodernism decline.⁴⁷ Good methods, good concepts, consistent vocabulary cannot be provided when the knowledge is not well developed and openly discussed and if the boundaries are still unclear or even completely unknown. These challenges cannot be faced by any research sector involved in the digital preservation without a contemporary double investment both in strengthening its own domain knowledge and competences and in supporting adequate and qualified cooperation processes.

Of course, many of the questions here proposed do not concern only the European institutions and, naturally, are not confined to the archival world. These demanding and global challenges involve the whole international scientific community at large, specifically scholars and professionals involved in the research and in the implementation actions. For this reason, it is more than probable that our academic and professional communities will not waste any possible present and future opportunity for strengthening their cooperation. But to reinforce their will and their reasoning capacities, it is important to recognize that not all the efforts already made are in the right direction and that it is essential to recover, with a stronger will and more intellectual consistency, the precious scholarly assets accumulated in our scientific studies and technical practices.

⁴⁶ “Elements for a Common Strategic Framework 2014 to 2020”, Accessed August 20, 2012, http://ec.europa.eu/regional_policy/what/future/index_en.cfm#2 and “Horizon 2020. The EU Framework Programme for Research and Innovation,” Accessed August 20, 2012, http://ec.europa.eu/research/horizon2020/index_en.cfm.

⁴⁷ E. Docx, Postmodernism is dead. In Prospect, 20 July 2011, <http://www.prospectmagazine.co.uk/magazine/postmodernism-is-dead-va-exhibition-age-of-authenticism>.

Models for National Collaboration

Coordination of the Digital Cultural Heritage in Sweden

Rolf Källman

The National Archives of Sweden

Abstract

On the basis of the national cultural policy, the European council conclusions, the responses from the heritage institutions and the digital agenda for Sweden - ICT for Everyone, the national strategy for digitization, digital access and digital preservation was established by the Swedish Government in December 2011. The aim of the strategy is to regulate the work for the government agencies and institutions that collect, preserve and make cultural heritage information available for the end users and furthermore to increase digitization and accessibility and to enhance digital preservation of digital cultural heritage holdings and collections online. In order to coordinate the continued development work on digitization issues, and coordinate the activities connected to the National Digital strategy, within the timeframe of 2012-2015, the government has established a coordinating secretariat for digitization, digital preservation and digital access to the cultural heritage – Digisam.

Author

Rolf Källman is Head of Department at The National Archives of Sweden, responsible for Digisam, the Swedish national secretariat for coordination of digitization, digital preservation and digital access to cultural heritage. Rolf has for more than 30 years worked with cultural heritage and heritage information in local, regional and national museums, and national authorities. For the last 15 years he has in various positions been engaged in the work with accessible and usable digital heritage information.

1. Introduction

For hundreds of years the cultural heritage information have been separated and handled in the frames of different academic disciplines. That has in many ways been a successful way to deepen the knowledge about our heritage, but in the perspective of the end users it has led to difficulties when it comes to search and retrieve complex cross domain information. With the rapidly growing amount of creative web based tools it has now become possible to promote access for everyone, researchers as well as pupils or the common interested citizen.

What happens when the collections and holdings held by archives, museums and libraries is digitized and how does it affect the daily work at our institutions when the digital born material rapidly grows? In what questions is it necessary to coordinate the work on a national level and work differently than in the past? And how can we make best use of powerful digital tools to gather, communicate and communicate our common heritage? These are some of the questions that Digisam—Swedish national coordination of digitization, digital preservation and digital access to cultural heritage has been commissioned by the government to manage.

In Europe, as in the rest of the world, digitization of our common cultural heritage is pointed out as a highly important business for knowledge building, amusement, democracy, innovation and economic growth.

This is clearly stressed in the European digital agenda, an important document within the framework of EU 2020, the European Union's strategy for economic growth, European Council Conclusions and ICT strategies on a national level.

The two most important Council Conclusions for the work with digitization of the cultural heritage is on Europeana, and on the digitization and online accessibility

of cultural material and digital preservation. It is here pointed out that it is essential to enable access for all to culture and knowledge in the digital era and to promote the richness and diversity of European cultural heritage which is an important resource for European cultural and creative industries, for contribution to economic growth and job creation and to the achievement of the digital single market through the increasing offer of new and innovative online products and services.

Common goals and broad cross boarder collaboration as a necessary platform is widely accepted among the member states. Europeana is established as the most important common platform and a strong and growing network. Europeana is the access point for European digital heritage and the natural environment for creative development and handling challenging questions that has to be solved along the line.

A necessary fundament for Europeana is a well organised technical infrastructure for digitization on national and regional levels.

The rapidly growing digitization is of course supported by the ongoing technical development, but to reach the goal of a free accessible and usable digital cultural heritage much more is needed

Digitization is a complex business that starts with strategic decisions and steering on a management level, and involves digital production with necessary steps taken for preservation and usability. The aims are of course clear defined effects in the society. This may sound trivial and commonly accepted, but when you look more closely into the single digitization projects on a national or a local level it becomes obvious that too many projects have not taken all the necessary steps into consideration with minor, to severe, negative consequences as a result and that lack of plans for long-term management and preservation jeopardizes tangible values and investments made.

The digital development gives the memory institutions and state media companies new opportunities to give citizens access to and use of cultural heritage. That is a strong statement in the Swedish national strategy for digitization, digital preservation and digital accessible cultural heritage materials and cultural information, the government has decided.

2. Towards coordination

In November 2009, the Swedish Government initiated the process aiming at a national strategy for digitization of our common cultural heritage.

A total of 27 cultural heritage institutions (archives, libraries and museums) under the rule of the ministry of culture and the ministry of education were assigned to deliver information and basic data for a strategy. Furthermore, 21 regional institutions responded to the invitation to contribute.

The responses varied a lot and were unfortunately difficult to compare with one another and it was therefore not possible to fully analyse and draw the picture of the situation as a whole. What was clear though was a lack of, but strong wish for, coordination and cross domain collaboration. It means that something has happened in the last 5-6 years. At that time, a common comment from memory institutions was that it was not relevant to coordinate their systems with other institutions because they had special and specific needs because since their collections were special and differed from other institutions. There

are several reasons for the changed attitudes and why the vast majority of the institutions today argue for the need for coordination on the basis of the lowest common denominators. Rising costs for IT management is one of the reasons, but primarily increased IT maturity and user requirements for seamless search and usability are the driving forces.

A lack of common terminologies was also obvious. The responses showed for example that the term digitization was used with different meaning, and was sometimes used for all kinds of IT-supported processes. The responses also showed a need for greater coordination within the field of standards, terminologies, infrastructure and long-term preservation of digital information.

3. A National Strategy for Digital Heritage

On the basis of the national cultural policy, the European council conclusions, the responses from the heritage institutions and the digital agenda for Sweden - ICT for Everyone, the national strategy for digitization, digital access and digital preservation was established by the Swedish Government in December 2011.

The aim of the strategy is to regulate the work for the government agencies and institutions that collect, preserve and make cultural heritage information available for the end users and furthermore to increase digitization and accessibility and to enhance digital preservation of digital cultural heritage holdings and collections online.

The strategy covers the whole range of digitization issues, from priorities to preservation and use. The focus is on the opportunities that digitization offers for use, reuse, participation and creativity.

Apart from fulfilling the commitment that follows with the conclusions in the European commission concerning Europeana, one of the government's cultural policy priorities, heritage for the future, underlines the importance of the digitization of our common cultural heritage. The Swedish culture policy also stresses the need for the government to continue efforts to digitize, preserve and disseminate cultural heritage

In the strategy it is decided that all government agencies and institutions covered by the strategy shall deliver plans for digitising and make their archives, collections and libraries available to the citizens. Digisam's role is to support them in their work by coordination and general guidance. The plans are stated to include selection criteria priorities, and the planned volume of digitized material is to be presented.

4. Digisam – a secretariat for National coordination of digitization, digital preservation and digital access to cultural heritage

In order to coordinate the continued development work on digitization issues, and coordinate the activities connected to the National Digital strategy, within the timeframe of 2012-2015, the government has established a coordinating secretariat for digitization, digital preservation and digital access to the cultural heritage—Digisam. Digisam started its work in the autumn of 2011 and is organized as a department at the National Archives. The main task is to promote the achievement of the objectives of the national strategy for digitization.

General guidelines, proposals for responsibility and how an integrated digital information management and a coordinated and cost-effective preservation should be designed are also key issues for Digisam to handle.

Digisam is therefore assigned to be a broad resource for questions about digitization, digital preservation and digital accessibility and will:

- Monitor and evaluate the institutions work within the task of the national strategy
- Work with issues of standards and systems for digital preservation,
- Conducting business intelligence, and collect and disseminate current research in the field of digitization and otherwise support the authorities and institutions
- Establish and arrange contacts with relevant organisations and private actors
- Work with and inform the participating authorities and institutions on the relevant issues of intellectual property rights
- Participate in working groups within the EU dealing with issues of digitization
- Explore possibilities for EU funding
- Provide support with communication, and seminar and training courses to participating agencies and institutions

It is important to underline that though Digisam is set up in the National Archives, its task is to be a resource for all state agencies, institutions and interested media companies in the field of cultural heritage.

Digisam's mission is focused on the state authorities and institutions, but the goal is that the results of the work will be beneficial and useful for anyone working with digitization and use of digital cultural heritage information. Therefore Digisam invite representatives from the private sector and civil society with an interest and involvement in digitization issues to participate in the work.

In the assignment to Digisam, the government has pointed out three key issues that are important to handle if the overall target is to be achieved. Those are common standards, priorities and usability

The government also has defined the deliverables as follows:

- Present recommendations for coordinated digital information management of collections and holdings.
- Develop proposals for cost-effective long-term digital preservation of collections and holdings.
- Define roles and responsibilities for the work on aggregation, access and preservation of digital cultural heritage information.

5. Collaboration between Digisam and the memory institutions

In the dialogue with the memory institutions, they accentuate steering and supporting guidelines for the selection of standards, formats, priorities, support for handling copyright issues, etc., as high on their wish list. The question that most institutions point out as the most difficult is the long-term preservation of digital information. Without secure and stable solutions for conservation, we cannot guarantee an accessible, use and re-use of heritage information neither today nor in 500 years. But how coordinated and cost-effective solutions for conservation to be designed are not obvious. The National Archives of course have a long experience within these matters which will be very useful in the continued work in collaboration with Digisam.

Preservation is just one of the issues where European cooperation is important. By participating in the joint work with Europeana and other projects and contexts, we can contribute to and benefit from a shared knowledge building and avoid reinventing the wheel, have extensive networks and funding opportunities. More Swedish heritage institutions need to be involved in European collaborative projects and Digisam hope to support such a development.

An important field of collaboration is of course the conversion process where analogue material is digitized. For various reasons it is sometimes necessary with small-scale and piecemeal digitization at a single institution, but in most cases there are considerable economic and rational benefits when digitization is made at large, more industrially organized production facilities, such as at MKC, the Swedish National Archives department for media conversion. On an average day about 100.000 digital is produced at MKC. *(not till <http://www.riksarkivet.se/default.aspx?id=19196&refid=1135>)*

Memory institutions have grown and evolved from their core of specific collections. . This means that the specificity of the collections has been the take-off point when systems and methods for managing have been chosen. When the digital tools were established at the institutions they came, in most cases only for use in the process of cataloguing and managing collections. It is therefore not difficult to understand that the systems were applied built to suit each institution's needs. There were not any collection systems to pick off the shelf.

On the road to a coordinated work within the field of digitization, preservation and use of digital information, the most commons needs that is highlighted by the institutions are common definitions and terminologies, common authorities and knowledge about the semantic web, support in IPR issues, long-term preservation, infrastructure and recommendations about almost everything.

6. How do we get there?

The list of all questions that need to be addressed in the context of Digisam's mission is long. We have a large and exciting task ahead of us to operate, collaborate, coordinate, inspire and be inspired. The only thing we know for sure is that it is essential to work close with all inside and outside the field of cultural heritage, who share our passion for these issues if we are to achieve the common objectives of the Swedish national strategy for digitization, preservation, accessible and usable digital heritage.

To achieve the objectives of increased digitization, preservation and use within established financial framework requires coordination and a clear allocation of responsibilities and roles. The strategy places the responsibility on the state authorities and institutions, but in order to reach the goals actors among the whole field of culture heritage institutions, public as well as private, and their collaboration partners, must be invited and involved in the work. A living used and re-used cultural heritage, today and tomorrow, is a concern of the whole society and therefore many voices must be heard.

On the highest level in the conceptual model we have defined the concepts steer, produce, use and preserve as central. This may seem obvious, but when look at the digitization projects that has been undertaken so far, we note that very few have taken all these four concepts in consideration. Very few of the digitization projects have started as an outcome of strategic plans, decided by the managements of the institutions. The institutions often holds a lot of knowledge and experience within the field of converting analogue data to digital, but too often the applied processes lead to either use or preservation and seldom take both these necessary outcomes in due consideration.

If Digisam reaches the goal of our mission, we by the year of 2015 have established the conditions, in a wide range of everything from infrastructure to practical advices for a successful work, for all actors involved in the complex issues of digitization. It is the archives, libraries, museums, other institutions and engaged citizens, with different missions and interests, that do the hard work making the information on our common cultural heritage accessible, usable and used. And it is they, the producers and users, that have the best prerequisites' to develop best use and benefits from new and old tracks. If Digisam shall be

able to support this, we need help from and to collaborate close with those who on different levels and with different perspectives are involved in the work.

7. Strategies

The most important strategy right now is to gather the state authorities and institutions that are targeted by the national strategy for digitization. A starting point for that is the plans for their digitization activities that they all have been commissioned to write. It is important that the plans will be powerful and effective tools for the single institution's work. And it is very important that the individual plans are interrelated. Digisam has offered the institutions to lead, coordinate and support the work with the plans. An important task in this process is to support the heads of the authorities and institutions to choose and define the necessary strategic decisions that has to be decided by the managements.

Interrelated plans as a common ground and consensus among the heads of the institutions, over the most strategic issues, will be a powerful tool in the work to come.

8. Challenges

Perhaps the greatest challenge of all is to think different. The technical development is today much faster than the development of creative use of information. And in the shift of perspectives and thinking it is important to be aware of not to, without reflection, copy an analogue way to work when it is transformed to a digital one. To look at the single institutions information as a part of a worldwide infrastructure is of outmost importance.

So there is definitely a need for a shift of focus. The building blocks in the knowledge processes are the data kept by the institutions. We have to accelerate our work with improvement of data quality with interoperability and sustainability as two of many important key words. A growing amount of linked open data leads to a necessity to review strategies and re allocate resources. The institutions need to see themselves as important nodes in a knowledge infrastructure. To raise quality of data is part of an ongoing information management, but the massive increase of large amounts of information is a resource-intensive work. Some data is also probably more important to improve the quality of than others, and here it is important to have clear strategies.

A strategy is to open up so that many more can join the work. And with the shift of focus follows a need for let go of control and invite other actors in the processes. Crowd sourcing is already a valuable tool, and the importance of and interest for it will probably accelerate sharply. The institutions have to play an important role as driving forces in the process to let in others and make good and innovative use of cultural heritage information that leads to surprisingly beneficial effects in the society.

9. The Users

On digitization follows that the public get a more open access to cultural heritage through the authorities, institutions and corporate archives and collections. It will also create more and new opportunities for research. Several of our memory institutions are now active in social media like Facebook, Flickr and Twitter where they share and discuss issues related to their collections and archives.

In many ways it is an open question who the users will be in the future. It's reasonable to assume that we will interact deeper and more frequent with our present user groups. But users and usage patterns change over time and with new open interfaces and possibilities to co-create we are likely to meet new user groups in situations that differ from the public areas we interact in today. What is definitely clear is that only imagination is the limit for the use of digital cultural heritage.

To a large extent the fact is that you ask for what you already know is available. We who works with cultural heritage information know that our institutions hold fantastically rich material that would be frequent demanded for if it was visible and accessible.

In this respect, students and teachers are most important user groups. Since cultural heritage information is an imprint of human activity in all its widths and depths, there are a rich potential for making use of heritage information in class rooms and learning tools. And for that a critical mass of information has to be made available as linked open data that can be used and put together in complex images and context in order to surprise us and create use and benefit where we least expect. And once again, to achieve that, long-term preservation is the key to and mother of used and re-used information.

Building and Preserving Library Digital Collections Through Community Collaboration

Victoria Reich

Executive Director, LOCKSS Program, Stanford University Libraries, USA

Abstract

Traditionally, libraries in the paper world were responsible for the safekeeping of much of society's cultural memory. They did this by building local collections, primarily to provide current local readers with rapid, convenient access to content. However, an important side effect of these local collections was a robust system for preserving the printed, and paper-based record for future readers. There were many copies of the content, on a durable, tamper-evident medium, under many separate administrations, with catalog systems capable of directing a reader to a nearby copy. Over the last two decades, the transition from paper-based publishing to Web publishing changed models for access and preservation for libraries. Instead of purchasing a copy, libraries now lease access to the publisher's copy. This has made preserving the contemporary published record more difficult, in these ways: Unlike paper, the publisher maintains the only legal long-term copies, and the publisher can alter or destroy them at will; Copyright law makes it problematic, if not impossible, to maintain other copies without specific permission from the publisher; Unlike paper, the copies are on a fragile, easily alterable medium with a short service life (magnetic disk or tape); Unlike paper, the format of the copies is expected to become rapidly obsolete. Preparations for this eventuality are expensive. The paper reviews these challenges, describing how the collaboration between libraries underlying the LOCKSS program enables each challenge to be addressed successfully.

Author

Victoria Reich is Executive Director of the LOCKSS Program, Stanford University Library, [www.lockss.org]. LOCKSS empowers libraries to locally archive and control their collections, ensuring perpetual access, local control and ownership, and the health and strength of the library community. Over 500 publishers participate in LOCKSS. Victoria helped to launch the CLOCKSS Archive and HighWire Press. She has extensive library experience, having held positions at Stanford University Libraries, the National Agricultural Library, the Library of Congress, and the University of Michigan. Victoria is on several Advisory Boards, including "Stanford Copyright & Fair Use". She received the 2008 Ulrich's Serials Librarianship Award.

1. Print Libraries as Preservation Systems

*"...the preservation of information resources is central to libraries and librarianship,"
and is a core value of our profession*

ALA, Preservation Policy 2001

In the paper world libraries are responsible for the safekeeping of much of society's cultural memory. They do this by building local collections, primarily to provide current local readers with rapid, convenient access to content. An important side effect of these local collections is a robust system for preserving the printed and paper-based record for future readers (Rosenthal 2005).

Proceedings of The Memory of the World in the Digital Age: Digitization and Preservation. An international conference on permanent access to digital documentary heritage, 26-28 September 2012, Vancouver, British Columbia, Canada, edited by Luciana Duranti and Elizabeth Shaffer (UNESCO 2013).

We tend to think of paper libraries as individual, stand-alone institutions. But if we think of the world's paper libraries as a network or distributed system of libraries, a different picture emerges. Viewed as a global system, libraries have attributes critical to their mission of preserving cultural memory that are not possessed by individual libraries. The important attributes of these paper library systems are:

- Highly replicated works have a better chance of surviving over many generations. The number of copies held by libraries reflects the importance of a work to various communities. More libraries hold the most important works, or most popular works implementing the meme; “lots of copies keep stuff safe”.
- Copies in libraries are distributed. The content is held by many libraries around the world, each under separate administrations, in countries with different laws, norms and customs. Some of these copies might be damaged by hurricanes, others by flood, earthquakes or fire. Others might be destroyed due to censorship, or accident. It is unlikely that all or even most copies of important works will be entirely destroyed, lost, or de-accessioned.
- Paper is a “tamper-evident” medium. It takes great skill to alter a printed book or journal in a way that does not provide obvious visual cues to the reader. Organizing personal visits to all the world's libraries to alter or destroy some book or journal article is difficult and expensive.
- Paper is a durable medium (Library of Congress 2012). Cotton papers and alkaline papers can last indefinitely.
- Paper libraries have local custody and control of their materials. Access is not dependent on continued payment and for as long as a library holds the title; their local clientele is guaranteed perpetual access. The purpose of preservation is to ensure content can be accessed.
- Libraries purchase and receive exactly what the publisher published. This is particularly true of content that's considered part of a library's “core collections”.

Despite the long and productive history of systematic cooperation among libraries, (an example is inter-library loan), the attributes described above that make the paper library system robust and effective in its role of preserving cultural memory were not deliberately created. Libraries did not have policies that mandated they cooperate to hold many replicas of publisher's content, in a highly distributed network, on a durable medium.

A primary motivation for paper library collections was to provide easy access for readers; an important business outcome of this practice is that paper libraries keep what they buy. The resulting robust preservation system of replicated copies was a fortunate accident enabling them to fulfill their charge to preserve access to the scholarly record for future generations.

2. Moving Safely from Print to Web

Over the last two decades, libraries have transitioned from acquiring paper-based collections to primarily leasing access to web-published ejournals and ebooks. The shift from owning paper collections to renting digital collections has had several disturbing unintended consequences. An unfortunate, simplistic implementation of web technology put libraries in a precarious and weakened position. Most libraries in the web environment are not fulfilling a traditional core social value; preserving their core collections for their future readers.

The move from paper to web has made preserving the contemporary published record more difficult in these ways:

- Most content is held by centralized organizations with limited redundancy and geo-political diversity.
- Copyright law makes it problematic, if not impossible, to maintain copies without specific permission from the publisher.
- Unlike paper, it is easy to alter digital copies with no trace or evidence.
- Unlike paper, the digital copies are on fragile, easily alterable medium with a short service life (magnetic disk or tape).
- Unlike paper, digital file formats are expected to become rapidly obsolete and many approaches to prepare for this eventuality are expensive (Kejser 2011).

3. Restoring Library Collection Infrastructure

The LOCKSS program (Lots Of Copies Keep Stuff Safe) a trademark of Stanford University, shows that a collaborative approach to the economic, legal, technical and social challenges of building and preserving library collections of digital content is both practical and economically sustainable. The program, founded in 1998, has low overhead and flexible commitments. The “Red Hat” model of free, open source software and paid support from libraries worldwide that use the system has kept the program in the black for more than five years with no grant funding.

In the web environment, most libraries rent access to electronic content. Leasing access to content puts libraries at risk; it outsources their duties as information stewards to third parties. A unilateral change of policy by the publisher or third party provider, or a failure to renew a subscription, can result in loss of electronic access to past material.

The LOCKSS Program, led and supported by libraries, gives libraries a way to preserve and control their own collections. Librarians use their local LOCKSS box to take custody of subscription and open access ejournals, ebooks, digitized collections, government documents, etc. A LOCKSS box is a library’s “digital stacks”, analogous to a library using its own buildings, shelves and staff to obtain, preserve and provide access to paper materials. The LOCKSS model restores libraries’ ability to build local collections, bringing the traditional purchase-and-own model to electronic materials and strengthening the library’s role in the digital age.

When libraries take custody of content to which they subscribe, access is separated from payment and perpetual access assured. Material stored in a local LOCKSS box remains available to members of the library’s local community even when the publisher goes away due to merger, bankruptcy, subscription cancellation, network outage or for any other reason. The content is always available to the local community, directly from the library, with no need to rely upon third parties. Locally owned collections guarantee 100% post cancellation access.

Copyright law makes it problematic, if not impossible, to maintain copies without specific permission from the publisher. Obtaining permission would be prohibitively expensive for libraries to do individually. When a publisher gives permission for the LOCKSS program to preserve their content, they give permission for all libraries that have authorized access to the content. There are many examples of a

single library negotiating with a publisher for LOCKSS preservation, to the benefit of all libraries in the LOCKSS network. As with paper libraries, the collaboration is implicit, not explicit.

Paper library collections consist primarily of materials received directly from the publisher. The libraries are providing to the readers what the publisher published. LOCKSS preserves what the readers sees on the web. The LOCKSS preservation approach is unique in delivering the publisher's original article, with the publisher's original branding and imprimatur. LOCKSS technology preserves the publisher's original content as of the date of web publication. The "look and feel" of the content, along with the publisher's branding, is preserved, resulting in an authentic representation of the authoritative source file. Readers see and use the most trusted, authoritative and contextually accurate version of important cultural materials.

Because the permission to collect and preserve the content is automatically granted to all authorized libraries, there are many libraries to share the cost of the technical work to get the content from the publisher. Again, as with paper libraries, the collaboration is implicit, not explicit.

The LOCKSS Program approach re-implements the key attributes of the paper library system, making it robust and effective at long-term preservation of our cultural heritage. A bit of cooperation among libraries that choose to take leadership and control has transformed seemingly intractable problems into successful solutions. Libraries take custody of content; they maintain local control of collections and assets. They keep what they buy and fulfill their duty to preserve access to the scholarly record for future generations.

4. Guarding Against Known Threats

In the paper world, many libraries each get a copy of a book. It is hard for a book to be accidentally or purposely destroyed because these many libraries are geographically dispersed, they are operated by many different administrations, under diverse political and cultural conditions. A primary motivation for paper library collections is to provide easy access for readers, however an important business outcome of this practice is that paper libraries keep what they buy. When considered as a network, these independent paper library collections resist damage; there is no central point of failure. Paper libraries fulfill their charge to preserve access to the scholarly record for future generations.

For digital materials, evidence suggests that human factors (intentional and unintentional) are the greatest cause of loss or corruption (Rosenthal 2011). Technology failures, economic failures and social failures also pose threats to the protection of digital content. When content is held in a single centralized repository, it is easy for someone to tamper with a master copy without detection. A dark archive, into which content disappears, only to reappear in a future emergency, does not engender confidence in either its availability or its correctness.

The LOCKSS network has similar tamper evident properties to print libraries. In the LOCKSS system, it is very difficult and expensive for someone to find and tamper with a significant number of the preserved copies without being caught. The copies are geographically distributed and independently held under many different administrations. Tamper-evidence engineering is a unique property of LOCKSS preservation and it is a keystone of our work (Rosenthal et al. 2005).

The LOCKSS Program is the only preservation approach that mitigates against the broad set of technical, economic and social threats to the security and long-term preservation of digital content. LOCKSS' ACM award-winning open-source technology is built on a peer-to-peer software infrastructure (Maniatis 2003) (Stanford 2012e).

It is unwise to trust closed source preservation systems; closed source systems are written and designed by employees operating under nondisclosure agreements. This constricted review limits opportunities to reveal software weaknesses and vulnerabilities that may interfere with preservation processes. Open source software supports cooperative, collaborative preservation systems in the following ways:

- Open source software is in itself a collaborative enterprise.
- Open source software is itself much better preserved, and much less likely to go obsolete, than closed-source software among other reasons because its copyright license terms encourage “lots of copies keep stuff safe” (Rosenthal 2009).
- Collaboration requires a degree of mutual trust, which is hard to sustain if no one can be sure what software anyone else is running. Open source software is transparent, in that anyone can read and experiment with the code.
- Open source promotes the use of open standards, and thus interoperability.
- Open source prevents vendor lock-in and thus contributes to keeping the cost of preservation as low as possible.

The LOCKSS peer-to-peer infrastructure requires that each library’s LOCKSS box cooperate with other LOCKSS boxes in the system to determine if content being preserved is what was originally collected from the publisher. If there are differences, the system repairs the content back to the authoritative version. The LOCKSS preservation processes employ implicit community collaboration.

The Blue Ribbon Task Force on Sustainable Digital Preservation and Access identified economics as the major threat to preservation (Blue Ribbon 2010). Even with optimistic projections of future costs, there is not enough money to preserve everything that should be preserved (Rosenthal 2012). This places particular importance on minimizing the cost of digital preservation systems, which has been a major focus of the LOCKSS Program since it was founded in 1998. Every unnecessary dollar of cost means more content that will be lost.

The community collaborates and shares responsibility for reaching out to publishers. The software is open source and has been contributed to by many in the community over time.

Multiple copies of content at different libraries audit each other and repair any damaged or missing content, eliminating the need for costly back-ups and manual auditing processes. The total system cost never appears on any one single budget and is thus never at risk from a single red pencil.

5. Cooperating for Robust Infrastructure

The LOCKSS Program builds your institution’s preservation infrastructure by re-implementing characteristics of the paper library system that enable libraries to keep society’s cultural memory safe in the digital environment. Librarians use the LOCKSS system to preserve content and to build their library’s local collections in two distinct networking environments, the Global LOCKSS Network and Private LOCKSS Networks. Local collections give libraries control over perpetual access for current readers and provide a robust preservation approach for future readers.

Librarians participating in the Global LOCKSS Network build and preserve open access titles and subscription ejournals and ebooks from over 520 participating publishers (Stanford 2012b). The Global

LOCKSS Network provides the digital equivalent to a library's general collections. (As with paper collections the publishers hold the intellectual property or copyright of these materials and libraries take custody of the materials for long-term safekeeping).

The Global LOCKSS Network leverages implicit cooperation among libraries worldwide. Sufficient replication is ensured because the materials preserved in the public network are those that the wider community has agreed they wish to preserve. The business relationship between the library and the publisher in the Global LOCKSS Network mirrors the print environment.

Private LOCKSS Networks are the digital equivalent of a library's special collections. The vast range of content genres preserved in Private LOCKSS Networks includes photo image collections, audio collections, government documents and databases. (For the most part, the libraries own the Intellectual Property or the copyright of the materials preserved in Private LOCKSS Networks).

Private LOCKSS Networks leverage explicit cooperation among a community of libraries that have common interests in specialized subject areas. These networks are highly targeted collaborative efforts among like-minded institutions sharing the preservation responsibility (including governance and sustainability) of e-content important to their institutions. Below are brief descriptions of several Private LOCKSS Networks, offered here as examples:

- Alabama Digital Preservation Network: Alabama libraries are collaborating to preserve a wide variety of historic archival materials, including image collections and databases (ADPN 2012).
- CLOCKSS Archive: Subscription and open access books, journals, and data are preserved in a network, spanning Europe, Asia and North America. When preserved content is no longer available from a publisher it is copied from the CLOCKSS Archive and made available for free, to everyone (CLOCKSS 2012).
- Council of Prairie and Pacific University Libraries: Consortium Canadian University libraries are collaborating to preserve collections important to the provinces of British Columbia, Alberta, Saskatchewan and Manitoba. This group has a particular focus on freely available born digital Web content including government documents, e-journals and small presses (COPPUL 2012).
- Data Preservation Alliance for the Social Sciences: Universities are collaborating to preserve social science data which include: opinion polls, voting records, surveys on family growth and income, social network data, government statistics and indices, and GIS data measuring human activity (Data-PASS 2012).
- Digital Federal Depository Library Program: Libraries in partnership with the U.S. Government Printing Office are working to ensure the distributed nature of the Federal Digital Library Program persists in the digital environment (Digital Federal Depository Libraries 2012).
- MetaArchive: Cooperative Run by the non-profit Educopia Institute, this international membership organization coordinates cultural memory organizations that are collaborating to preserve very high value locally created digital materials (MetaArchive 2012).

6. Building Distributed Digital Preservation Networks

The LOCKSS system was the world's first production quality Distributed Digital Preservation Network. Distributed Digital Preservation Networks are designed with intentional geographical and organizational

distributed infrastructures as essential components of their preservation models. The LOCKSS Program set best practices for reliable preservation and persistent access to digital content via content replication, geographic distribution, infrastructure heterogeneity, modularity and organizational diversity.

Distributed Digital Preservation networks require community collaboration and cooperation. It is not possible to achieve the required robustness to ensure the long-term persistence of digital objects through centralized technical or organizational approaches. Beware organizations that claim distributed digital preservation infrastructure when their technical implementation is a few back-up copies around the globe.

Implementing the LOCKSS Distributed Digital Preservation approach requires three actions: a publisher to give permission for the target content to be preserved; for a library to bring online a LOCKSS box that has authorized access to the content; and for that LOCKSS box to be registered with one of a number of associated LOCKSS Alliance networks (Stanford 2012d).

Publishers grant the LOCKSS system legal permission to ingest, preserve and access their intellectual content by putting online a LOCKSS permission statement. The LOCKSS permission statement is bundled with the content so that the content and the legal rights are preserved together. Paper contracts are hard to track through time, preserving the legal agreement with the contract minimizes any future misunderstandings.

The Global LOCKSS Network implements an extremely effective form of cooperative collection development. When one library or library group secures a publisher's agreement to participate in LOCKSS, all LOCKSS libraries with authoritative access to that content have permission to locally ingest and preserve the material.

A library uses the freely available, open source LOCKSS software to turn a mid-range PC, or virtual computing environment into a digital preservation appliance called a LOCKSS box (Stanford 2012c).

A LOCKSS box performs five main functions:

- It **ingests** content from target websites using a web crawler similar to those used by search engines.
- It **preserves** content by continually comparing the content it has collected with the same content collected by other LOCKSS boxes, and repairing any differences.
- It **delivers** authoritative content to readers by acting as a web proxy, cache or via Metadata resolvers when the publisher's website is not available.
- It provides **management** through a web interface that allows librarians to select new content for preservation, monitor the content being preserved and control access to the preserved content.
- It dynamically **migrates content** to new formats as needed for display.

LOCKSS Program staff at Stanford University analyses the target content's URL structure, file formats and delivery mechanisms. They design, implement and update a tailored, content-specific preservation action plan that serves publishers, librarians and readers.

The publisher permits the LOCKSS system to collect, preserve and provide access to the content by putting a LOCKSS manifest page on the content's website (Stanford 2012c). The manifest page contains a LOCKSS permission statement and links to the issues (or other parts) of the content as they are

published. The required manifest page is ingested and preserved with the original content and negates the need for paper contracts.

Software called a LOCKSS plug-in tells each institution's LOCKSS box where to find the publisher's LOCKSS manifest page, and how far to follow the chains of web links. A LOCKSS plug-in encapsulates a publisher's content model by listing parameters specific to each publishing platform. The LOCKSS team builds, tests and distributes plug-ins to LOCKSS boxes registered with the LOCKSS Alliance.

Every LOCKSS box is located at an IP address that falls within its parent University's IP address range. Authorized LOCKSS boxes independently collect subscription or open access content directly from the publisher's website. The publisher authorizes or denies a LOCKSS box's access to content through their access control system. Publishers register LOCKSS activity on their web logs and have access to real time statistics through their own systems.

Once ingest is complete, the LOCKSS technology ensures that each LOCKSS box has collected all intended content, thus preserving the authoritative version. The LOCKSS software continually monitors the content in each LOCKSS box to ensure it is properly preserved through a cooperative preservation process that compares one LOCKSS box's content with the same content on other LOCKSS boxes. When content is damaged or lost the system arranges for content repair from another LOCKSS box.

The administrator of each LOCKSS box can monitor the preservation status of the content in their box, by looking at delivered content and the management tools available through the LOCKSS box web administrative interface (Stanford 2012f).

7. Providing Continuous Access

An institution's LOCKSS box can provide readers with continual, seamless access to branded publisher content (Stanford 2012c). The LOCKSS system preserves content at its' original URL, critically retaining the content's relationship to other web resources. An institution's LOCKSS box delivers content to authorized readers only when the publisher's website is unavailable (subscription canceled, network traffic, publisher server down). The LOCKSS Program works to preserve and to deliver the publisher's original artefact to readers; in other words, what the publisher published.

LOCKSS boxes provide four main ways for readers to access the content they preserve: by proxying (i.e., acting like a web cache), by serving (acting like a web server) or by serving through integration with an OpenURL resolver, or the emerging Memento standard.

- **Proxying:** Institutions often run web proxies to allow off-campus users to access subscription content. Libraries that integrate their LOCKSS box into a proxy (PAC Files, EZ Proxy, ICP, Squid) ensure a reader's URL request is seamlessly fulfilled when the content is unavailable from the publisher's website.
- **Basic Serving:** In the basic serving model, articles are accessed using a local URL pointing to the LOCKSS box. The LOCKSS box checks if the publisher will provide content to fulfill a reader's request. If the content is not available from the publisher, the LOCKSS box serves its own copy to the reader.
- **OpenURL Serving:** Libraries can integrate their LOCKSS box with their library catalog and OpenURL resolver by adding their LOCKSS box as a target to an OpenURL Resolver.

- Memento: The Internet Engineering Task Force (IETF) is in the process of standardizing Memento, a mechanism created by Michael Nelson and Herbert van de Sompel by which browsers can access preserved versions of websites. With funding from the Mellon Foundation, the LOCKSS team is currently implementing Memento so that LOCKSS boxes will conform to this standard to access past content.

Three audit and verification tools detail what content is in a library's LOCKSS box and the content's preservation status.

- On demand, a LOCKSS box produces a KBART (Knowledge Bases And Related Tools) report of the locally preserved content.
- A LOCKSS box displays detailed preservation status for each Archival Unit. (An Archival Unit is typically a volume of a journal, or a complete book).
- A LOCKSS box administrator can use a properly configured web browser from an authorized IP address to view preserved content through an "audit proxy". The viewer sees the content as it was collected by the LOCKSS system.

Librarians administer their institution's LOCKSS boxes through a web browser that allows them to easily select new content for preservation, monitor content's preservation status and a variety of other functions (Stanford 2012c).

Post cancellation access to all preserved content is ensured as the content is under the library's local custody.

8. Migrating Obsolete Formats

LOCKSS preserves all web published formats (animations, datasets, moving images, still images, software, sound, text) and genres (journals, books, blogs, websites, scanned files, audio, video). The LOCKSS software is format-agnostic and preserves all content in its original format, as delivered from the publisher, including the format metadata that enables a browser to render the content.

The field of digital preservation has lavished much attention on the risk of format obsolescence (Rosenthal 2009). For web content, a format becomes obsolete when commonly available browsers and plug-ins can no longer render it. There is little evidence that this is happening to the widely used web formats in which the books and journals of the GLN or the special collections in PLNs are published.

Nevertheless, more than seven years ago the LOCKSS technology demonstrated its ability to handle format obsolescence if and when it occurs (Rosenthal et al. 2005). When a browser requests content from a LOCKSS box, it uses the part of the HTTP standard called "content negotiation" to specify the formats it can render. If the requesting browser cannot render the format of the preserved content, the LOCKSS box invokes appropriate format migrators to create a temporary copy in a format the browser can render (Ockerbloom 1998). After use, the LOCKSS box discards the temporary access copy.

(The LOCKSS team demonstrated this process by modifying a browser to claim that it could not render the GIF format, whereupon the LOCKSS box created temporary copies of the GIF images in a histology paper in PNG format. An unmodified browser saw the same images as GIFs).

The LOCKSS Program's "migration on access" approach has significant advantages over "format normalization" as it preserves the original artefact, uses much less overhead, saves money and takes advantage of the most up to date technology. Preserving the content in its original format satisfies

archival requirements. It allows the LOCKSS system to be frugal with storage space. We know of no preservation system that discards the original bits after migrating them to a new format. Migrating and keeping both the original and the migrated copy multiplies the storage requirements for a preservation system by the number of migrations.

Preserved content is migrated by the most recent, and presumably best, technology available at the time the reader requests access. Preserved content is rarely accessed. Performing migration only when and if it is needed reduces the resource cost. Content can be migrated directly from the original to the current format, minimizing the effects of format conversion artefacts. The format converters, once developed, can themselves be preserved to document the original format.

9. Taking Action

Paper library systems, through implicit library collaboration and cooperation are robust and effective at preserving cultural memory. Libraries have an opportunity and a responsibility to acquire these characteristics in the digital environment.

The Stanford University LOCKSS Program provides libraries with the means to cooperate socially, technically and financially to build and preserve authoritative copies of subscription, open access and special digital collections. The open source distributed digital preservation approach builds tamper evident infrastructure. The time-tested robust and resilient paper world implementation of “lots of copies keep stuff safe” is easily implemented in the digital environment. Local infrastructure and collections ensure continual access and the more content is replicated, the greater its chances of surviving for the long-term. Collaborative, collective action at a local level yields highly leveraged results globally.

References

- ADPN - Alabama Digital Preservation Network. 2012. Accessed August 22, 2012. <http://www.adpn.org/>.
- ALA - American Library Association. “Core Values of Librarianship.” Accessed August 19, 2012. <http://www.ala.org/offices/oif/statementspols/corevaluesstatement/corevalues>.
- Blue Ribbon Task Force on Sustainable Digital Preservation and Access. “Sustainable Economics for a Digital Planet - Ensuring Long-Term Access to Digital Information.” Final Report, February 2010. Accessed August 19, 2012. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.
- CLOCKSS Archive. 2012. Accessed August 22, 2012. <http://www.clockss.org>.
- Council of Prairie and Pacific University Libraries: Consortium Canadian University libraries (COPPUL).2012. Accessed August 22, 2012. <http://www.coppul.ca/>.
- Data-PASS - Data Preservation Alliance for the Social Sciences. 2012. Accessed August 22, 2012. <http://www.data-pass.org/>.
- Digital Federal Depository Libraries. 2012. Accessed August 22, 2012. <http://www.lockss.org/community/networks/digital-federal-depository-library-program/>.
- Kejser, Ulla Bøgvad, Anders Bo Nielsen and Alex Thirifays. “Cost Model for Digital Preservation: Cost of Digital Migration.” *The International Journal of Digital Curation* 6, no. 1 (2011): 255-267. Accessed August 22, 2012.. <http://www.ijdc.net/index.php/ijdc/article/download/177/246>.

- Library of Congress. "The Deterioration and Preservation of Paper: Some Essential Facts." 2012. Accessed August 19, 2012. <http://www.loc.gov/preservation/resources/care/deterioratebrochure.html>.
- Maniatis, Petros, David S. H. Rosenthal, Mema Roussopoulos, Mary Baker, TJ Giuli, and Yanto Muliadi. "Preserving Peer Replicas By Rate-Limited Sampled Voting." In *Proceedings of SOSP '03, October 19-22, 2003, Bolton Landing, New York*. (Best paper presented at the 19th ACM Symposium on Operating Systems Principles (SOSP)). ACM, 2003. Accessed August 19, 2012. <http://www.eecs.harvard.edu/~mema/publications/SOSP2003-long.pdf>.
- MetaArchive. 2012. Accessed August 22, 2012. <http://www.metaarchive.org>.
- Ockerbloom, John. "Mediating Among Diverse Data Formats." Ph.D. diss., Carnegie Mellon Computer Science Department, Pittsburgh, 1998. Accessed August 19, 2012. <https://www.cs.cmu.edu/~spok/thesis.ps>.
- Rosenthal, David S. H. "How Few Copies?" Paper presented at Screening the Future 2011, Netherlands Institute for Sound and Vision, March 14 -15, 2011. Accessed August 19, 2012. <http://blog.dshr.org/2011/03/how-few-copies.html>.
- Rosenthal, David S.H. "How Well Are We Ensuring the Longevity of Digital Documents?" Keynote paper presented at the CNI Spring Task Force Meeting, April 6-7, 2009, Minneapolis MN. Accessed August 19, 2012. <http://blog.dshr.org/2009/04/spring-cni-plenary-remix.html>.
- Rosenthal, David S. H. "Modeling the Economics of Long-Term Storage." Paper presented at the Personal Digital Archiving Conference, Internet Archive, San Francisco. February 22-24, 2012. Accessed August 19, 2012. <http://blog.dshr.org/2012/02/talk-at-pda2012.html>.
- Rosenthal, David S. H., Thomas Robertson, Tom Lipkis, Vicky Reich, and Seth Morabito. "Requirements for Digital Preservation Systems: A Bottom-Up Approach." *D-Lib Magazine* 11, no. 11 (November 2005). Accessed August 19, 2012. doi:10.1045/november2005-rosenthal.
- Stanford University LOCKSS Program. "Build A LOCKSS Box." 2012a. Accessed August 22, 2012. <http://www.lockss.org/support/build-a-lockss-box/>.
- Stanford University LOCKSS Program. "Global LOCKSS Network Publishers and Titles." 2012b . Accessed August 22, 2012. <http://www.lockss.org/community/publishers-titles-gln/>.
- Stanford University LOCKSS Program. "How LOCKSS Works." 2012c. Accessed August 22, 2012. <http://www.lockss.org/about/how-it-works/>.
- Stanford University LOCKSS Program. "How To Join." 2012d. Accessed August 22, 2012. <http://www.lockss.org/join/>.
- Stanford University LOCKSS Program. "Preservation Principles." 2012e. Accessed August 22, 2012. <http://www.lockss.org/support/use-a-lockss-box/>.
- Stanford University LOCKSS Program. "Use A LOCKSS Box." 2012f. Accessed August 22, 2012. <http://www.lockss.org/support/use-a-lockss-box/>.

National Library of New Zealand, Digital Preservation and the Role of UNESCO

Steve Knight

Programme Director Preservation Research and Consultancy, National Library of New Zealand

Abstract

In 2003, UNESCO published its Guidelines for the Preservation of Digital Heritage and adopted its Charter on the Preservation of Digital Heritage. Also in 2003, The National Library of New Zealand Act was passed providing the mandate for the Library to begin collecting and preserving New Zealand's digital heritage in ways that will ensure current and future access. This paper will show that the 2003 UNESCO documents are still valid today, look briefly at the current state of digital preservation and outline some possible approaches for UNESCO to refresh its 2003 statements on digital preservation.

Author

Steve Knight is the Programme Director, Preservation Research and Consultancy at the National Library of New Zealand (NLNZ). From a library background Steve has experience in a range of information management disciplines, including records management and document management and the design and implementation of electronic services. Among other initiatives Steve was involved in setting up the National Digital Forum in New Zealand, the award-winning Matapihi collaboration and the National Digital Heritage Archive, NLNZ's digital preservation programme. Steve has represented the National Library of New Zealand on the National Digital Forum Board, is a member of the Advisory Committee for the School of Information Management at Victoria University in Wellington, is on the Steering Committee for the International Internet Preservation Consortium and has represented the National Library in various European Commission activities related to the Commission's work on digital libraries and digital preservation.

UNESCO Charter on the Preservation of Digital Heritage 2003.	National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003
<p>The General Conference</p> <p>Considering that the disappearance of heritage in whatever form constitutes an impoverishment of the heritage of all nations,</p> <p>Recognising that such resources of information and creative expression are increasingly produced, distributed, accessed and maintained in digital form, creating a new legacy—the digital heritage,</p> <p>Understanding that this digital heritage is at risk of being lost and that its preservation for the benefit of present and future generations is an urgent issue of worldwide concern,</p> <p>Proclaims the following principles and adopts the present Charter.</p>	<p>The purpose of the National Library is to enrich the cultural and economic life of New Zealand and its interchanges with other nations by, as appropriate, collecting, preserving, and protecting documents, particularly those relating to New Zealand, and making them accessible for all the people of New Zealand, in a manner consistent with their status as documentary heritage and taonga;</p> <p>For the purposes of carrying out his or her duties, the National Librarian and any employee, contractor, or agent of the chief executive may possess, copy, store in electronic form (whether offline or online), and use any copy of a deposited document.</p>

1. Introduction

At an almost identical point in time UNESCO and the National Library of New Zealand (NLNZ) formally recognise the importance of digital preservation to the safety and protection of the world's digital heritage, UNESCO through the Charter on the Preservation of Digital Heritage¹ and NLNZ through the revision of its legislation which provided the mandate for the Library to collect in the digital realm and which catalysed the Library's digital preservation programme.²

Earlier in 2003, UNESCO had released its Guidelines for the Preservation of Digital Heritage prepared by Colin Webb, then Director of Preservation at the National Library of Australia.³ In 2004 the Library launches its National Digital Heritage Archive (NDHA) project, the foundation for its digital preservation programme. The project is funded by government to the tune of NZ\$24 million and has a four year timeframe.

In this paper, I will undertake to do three things:

1. Show the current validity of the 2003 UNESCO documents with specific reference to the statements on Responsibility and Principles
2. Briefly comment on the current digital preservation environment
3. Outline some possible approaches for UNESCO to refresh its commitment to the digital preservation domain and take a leading role in the long-term preservation of the digital heritage of all nations.

The point here is not to show convergence between UNESCO conceptualising as expressed in the Charter and Guidelines and evolving practice at the NLNZ. The point is to show how robust and comprehensive this early work from UNESCO was and how beneficial that work still is to NLNZ, as a representative organisation implementing a digital preservation programme.

2. Responsibility for digital preservation

In its discussion of responsibility the UNESCO Guidelines pose four questions to help determine an organisation's for accepting a digital preservation responsibility:

1. Does the business of the organisation imply an existing or potential preservation obligation for any kinds of digital heritage materials? (Is the organisation required to take responsibility?)
2. Does the organisation have an interest in accepting a preservation responsibility? (Does it want to have a role?)
3. Does the organisation have, or could it acquire, the capacity to take on a preservation responsibility?
4. Is this really someone else's responsibility?

¹ UNESCO, "Charter on the Preservation of Digital Heritage," 2003, accessed August 13, 2012, http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html.

² National Library of New Zealand, *National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003*, accessed August 13, 2012, <http://www.legislation.govt.nz/act/public/2003/0019/latest/DLM191962.html>.

³ UNESCO, "Guidelines for the Preservation of Digital Heritage," 2003, accessed August 13, 2012, <http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>.

For the National Library of New Zealand the answers to these questions are unequivocally yes, yes, yes and no. There is a natural fit between a national library and the need to ensure the long-term preservation of a nation's digital heritage.

The Guidelines also propose some possible continuums for levels of responsibility an organisation might agree to undertake.

<i>1. Scope of material</i>	Restricted Programme			Selective Programme	Broad Programme
	very restricted				wide range, comprehensively collected
<i>2. Scope of time</i>	Initial Programme	Caretaker Programme			Long-term Programme
	only until technology changes	only until use ceases	for a limited number of years		"forever"
<i>3. Scope of functions and responsibilities</i>	Partial, non-comprehensive Programme				Comprehensive Programme
	restricted functions				comprehensive functions
<i>4. Level of reliability</i>	Non-reliable Programme				Fully reliable Programme
	limited characteristics of reliability				all characteristics of reliability

Using these proposed continuums of responsibility the Library's responsibilities for the digital heritage of the nation requires a broad programme, a long-term programme, a comprehensive programme and a fully reliable programme.

The National Library of New Zealand cannot avoid digital preservation.

3. Principles of digital preservation

The following table reflects the NLNZ approach to the principles of digital preservation articulated in the UNESCO Guidelines.

Twenty-two of the forty-one principles are reflected upon here. The reason a subset of the principles has been chosen is that one principle is often refined further by another principle, e.g.,

maintaining accessibility. The purpose of this exercise is to show the overall alignment of the UNESCO Principles and the NLNZ digital preservation programme as it developed over time.

Number	UNESCO Guidelines principle	NLNZ approach
1	Not all digital materials need to be kept, only those that are judged to have ongoing value: these form the digital heritage.	The goal of the NDHA programme is to ‘enable the National Library of New Zealand ... to collect, make accessible, and preserve in perpetuity, New Zealand’s digital heritage, as defined by the Library’s current collection policy’, a clear recognition that similar decisions need to be made when building digital collections as with analogue collections.
3	Digital materials cannot be said to be preserved if access is lost.	As above the goal of the NDHA programme was to ‘to collect, make accessible, and preserve in perpetuity’.
5	Digital preservation will only happen if organisations and individuals accept responsibility for it.	A national library is a natural fit for digital preservation. The National Library Act 2003 gave NLNZ the mandate to collect digital and the impetus to develop its digital preservation programme. The core issue here is to be able to transfer our trustworthiness from the analogue world to the digital world.
8	It is important to do no harm.	The library identified nine architectural qualities as being of particular significance to the NDHA. The highest ranked of these is Data Assurance expressed as ‘zero data loss.’ The complexity between the statement and the practical import of the statement is the challenge.
9	Acceptance of responsibility should be explicitly and responsibly declared, taking account of the likely implications for other preservation programmes and for other stakeholders.	Acceptance of responsibility is an explicit outcome of the National Library Act 2003. We are currently working with Archives New Zealand to leverage the NDHA to support the preservation of the digital public record in line with the Public Records Act 2005. We are also working with external organisations with regard to third party hosting and the potential for a whole-of-country approach to digital preservation.
14	Working with producers to influence the standards and practices they use, and to increase their awareness of preservation needs, are important investments.	The Library is very aware that there is a continuum of activities that influence the nature and extent of the digital preservation programme’s ability to undertake its role. We are beginning discussions with major publishers to ensure that they are integral partners in the broadening of the Library’s digital preservation programme.
16	Digital heritage materials must be moved to a safe place where they can be	The Library has its own dedicated data centre with discrete provision for the digital preservation programme. This is currently being enhanced with movement of the NDHA to a

	controlled, protected and managed for preservation.	state-of-the-art data centre located outside of the Library and built to withstand one-in-a-hundred-year events. This is necessary in a geologically active region.
17	Digital heritage materials must be uniquely identified, and described using appropriate metadata for resource discovery, management and preservation.	All objects in the NDHA receive a unique and persistent identifier (using the Handle System). ⁴ As well as being uniquely identified in this manner each item is discretely described in the Library resource discovery and collection management systems and data is automatically synchronised between the collection management and digital preservation systems.
19	Preservation programmes should use standardised metadata schemas as they become available, for interoperability between programmes.	The Library's resource discovery and collection management systems are based on MARC and ISAD-G metadata. Synchronisation between collection management and digital preservation systems is via Dublin Core. In 2003 the Library presented a revised version of its earlier work on preservation metadata ⁵ including an associated XML schema. ⁶ With the development of the Library's digital preservation system Rosetta, by Ex Libris Group, the underlying data model has been created in accordance with the de facto standard PREMIS. ⁷
20	The links between digital objects and their metadata must be securely maintained, and the metadata must be preserved.	The NDHA programme was developed within the framework provided by NASA's Open Archival Information System (OAIS) reference model. ⁸ The Library has core enterprise systems for resource discovery and collection management. These have now been joined by Rosetta, a purpose built digital preservation system. Together these provide the links between objects and their metadata required to ensure the long-term viability of the objects over time.
21	Authenticity is a critical issue where digital objects are used as evidence.	All objects ingested into the NDHA permanent repository receive a combined hash key of MD5, CRC32, and SHA1 for use in the detection of change in their files. Regular, ongoing crawling of the permanent repository is undertaken in order to

⁴ Corporation for National Research Initiatives, "Handle System," accessed August 18, 2012, <http://www.handle.net/>.

⁵ National Library of New Zealand, "Metadata Standards Framework – Preservation Metadata (Revised)," 2003, accessed August 13, 2012, <http://www.natlib.govt.nz/downloads/metascema-revised.pdf>.

⁶ National Library of New Zealand, "XML schema for the preservation metadata dictionary," 2003, accessed August 13, 2012, http://www.natlib.govt.nz/files/nlnz_presmet.xsd.

⁷ PREMIS Editorial Committee, "PREMIS Data Dictionary for Preservation Metadata," Version 2.2, 2012, accessed August 13, 2012, <http://www.loc.gov/standards/premis/>.

⁸ The Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, 2009, accessed August 18 2012, <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf>.

		detect change in the files stored there.
22	Data that underlies digital objects must be safely stored and managed if there is to be any chance of re-presenting authentic objects to users.	Issues regarding the preservation of databases and research data are still to be formally addressed by the NDHA. This is seen as a major gap in our current capability.
24	Authenticity is best protected by measures that ensure the integrity of the data is not compromised, and by documentation that maintains the clear identity of the material.	Our digital preservation system Rosetta currently contains data about 276 separate events that occur in the system. Of these fifty are considered provenance events and are retained with their relevant objects for the long-term.
25	Data protection is built on the principles of system security and redundancy.	The NDHA defines nine core architectural qualities: Data Assurance, Security, Portability, Flexibility, Manageability and Maintainability, Scalability, Performance, Availability, Disaster Recovery. We have not yet tested how well we have met our goals in terms of achieving these qualities.
28	Preservation action should not be delayed until a single 'digital preservation standard' appears.	<p>Migration and emulation are not yet well proven in the digital preservation domain, in particular the quality assurance processes for determining success of any particular transformation.</p> <p>Understanding of formats, tools for characterising, validating and extracting metadata from formats are all less satisfactory than we would like.</p> <p>However, these are not reasons for not moving towards a functioning digital preservation programme. The NDHA has material flagged as format 'unknown'. We expect to come back to this material when resources, tools, capability, etc., allow. In the meantime it is in a managed digital preservation environment.</p>
30	It is reasonable for programmes to choose multiple strategies for preserving access, especially to diverse collections. They should consider the potential benefits of maintaining the original data streams of materials as well as any	<p>The NDHA maintains bit streams, Preservation Masters, Modified Masters (e.g., sound recordings with noise removed) and access derivatives appropriate for the individual object.</p> <p>The hope is that we should be able to move seamlessly between levels of objects over time as new tools and technologies arise that enable better preservation and access processes for those objects.</p>

	modified versions, as an insurance against the failure of still uncertain strategies.	
32	Preservation programmes are often required to judge acceptable and unacceptable levels of loss, in terms of items, elements, and user needs.	Zero data loss is the goal. However, that may not be possible for all material at all times, if at all. Pre-conditioning (making objects preservation ready), preservation planning and action may require acceptance of loss. Within the NLNZ programme this is a curatorial decision based on the best advice possible from the digital preservation programme at the time.
37	The costs of preservation programmes are hard to estimate because they encompass so much uncertainty.	Issues regarding the costs of digital preservation at NLNZ are still to be formally addressed. The move to Infrastructure as a Service (IaaS) in the next 12 months will provide more certainty about the costs of storage in the first instance.
38	Preservation programmes may start as pilot projects but they eventually need to establish sustainable business models.	Issues relating to the business model for digital preservation have not yet been addressed by NLNZ. As we move to IaaS, 3 rd party hosting and the potential for a whole of country approach to digital preservation this will become increasing important.
39	While suitable service providers may be found to carry out some functions, ultimately responsibility for achieving preservation objectives rests with preservation programmes, and with those who oversee and resource them.	It is noteworthy that commercial service providers are conspicuous by their absence and where services are starting to appear they are in the 'softer' areas of the preservation domain, e.g., in the area of audit and certification, but not in the development of commercial emulation services. Is it that there is not a sustainable market for such services? If that is the case how does the community respond to this?
40	Working collaboratively is often a cost effective way to build preservation programmes with wide coverage, mutual support and the required expertise.	Rosetta is one of only two commercially available digital preservation systems in the world. NLNZ decided early on to work with commercial partners (Ex Libris and Sun Microsystems). We did not have the skills in-house to build a solution. We did not want to build a bespoke solution for the National Library of New Zealand. We believe our activities should be applicable as broadly as possible across the whole digital preservation community. Consequently, development of a broad-based user community, a formal programme of community agreed enhancements and a development roadmap among other things convinced us that the commercial route was more likely to provide the continuity that we needed.
41	Collaboration involves	As part of the NDHA project we instituted an international

	costs and choices as well as potential benefits.	Peer Review Group to provide a check on the Library's thinking about digital preservation and also to act as an objective advisor to the Library's vendor as to whether what we were asking for did actually represent current best practice thinking about what a digital preservation system should do. We also instituted a Cross Government Working Group from 27 government departments to help catalyse awareness of digital preservation and also to tap into specialist expertise from the wider public sector.
--	--	---

From this brief discussion of the responsibilities and principles delineated by the UNESCO Guidelines and how they have been responded to (albeit unwittingly) by the National Library of New Zealand, we can see the quality of the Guidelines approach and its robustness. It is worth considering in particular the foresight required to articulate some very difficult practical issues, e.g., Principles 8, 22, 25 and 28.

4. Digital Preservation

In 2007, the amount of digital information created surpassed, for the first time, the amount of storage needed to deal with it. Of course, we don't need to store all the bits created like digital TV signals or phone-call routing information. But if we wanted to, we couldn't.⁹

About the only growth rate that hasn't gone negative since the beginning of the recession in 2008 is the creation of new digital information. People are still taking pictures, making phone calls, sending emails, blogging, and putting up videos on YouTube. Enterprises are still capturing daily transaction records and adding to their data warehouses. Governments are still requiring more information be kept and protected, forcing the migration to digital TV and taking surveillance photos of their citizens.

But what is the current state of our ability to collect this digital deluge and preserve it for the long-term?

A quick perusal of the digital preservation world today shows over fifty initiatives working on digital preservation systems, repositories, projects and advisory programmes and the provision of standards, products, tools and services (and this from a decidedly narrow, monolingual/English perspective).

When we look at how we talk about digital preservation (repositories, data archiving, digital archiving, life cycle, digital curation, data curation digital preservation) and when we look at the social and cultural issues related to digital preservation (standards, audit/certification, technical, organisational, legal, economic, education) are we presented with a clear, coherent framework for the long-term preservation of nations' digital heritage?

It could be argued that there is currently a genuine lack of a clear, shared understanding of what digital preservation is, what it should be doing, what questions it should be answering for the future and

⁹ John F. Gantz, Christopher Chute, Alex Manfrediz, Stephen Minton, David Reinsel, Wolfgang Schlichting, and Anna Toncheva, "The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011," March 2008, An IDC White Paper, sponsored by EMC, accessed August 24, 2012, <http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>.

what sort of information/data we should be providing today to allow those questions to be accurately answered in the future.

This needs to be addressed, as it is digital preservation that will ensure maximum leverage of the digital long-tail. Governments everywhere are concerned about the cost to the public and the sustainability of government information. Government data, big data, linked data, open government, open access, citizens' rights to information, data re-use all permeate our discussions about digital information. Digital preservation needs to be at the centre of this discourse.

Concurrent with this however, as noted recently in *Wired Magazine*, open data is not just about empowering the empowered and open data is not an end in itself. 'Massive data dumps and even friendly online government portals are insufficient,' ... 'ordinary people need to know what information is available and they need the training to be conversant in it' ... 'and if people are to have more than theoretical access to the information, it needs to be easy and cheap to use.' 'That means investing in the kinds of organizations doing outreach, advocacy, and education in the communities least familiar with the benefits of data transparency. If we want truly open government, we still have to do the hard work of addressing basic and stubborn inequalities. However freely it flows, the data alone isn't enough.'¹⁰

In her keynote at the Aligning National Approaches to Digital Preservation conference in Tallinn last year Laura Campbell broached the possibility of 'an international preservation body with a focus on policy, perhaps assisted by an advisory expert group to identify what categories of digital objects are most at risk. The body could promote an international notion of collection, work on standards and tools, and maybe maintain a common index of preserved materials.' The results of that conference have recently been published with six key strategies for alignment described—legal, organizational, standards, technical, economic and education.¹¹

Maybe here lies the solution to the current fractured state of the digital preservation domain mentioned above, and maybe also a role for UNESCO.

5. Facets of the web

While the current explosion of digital materials as the result of global digitization programmes creates the need for a supporting digital preservation infrastructure this is not where the primary risk for digital resides. The real risk is embodied in digitally born materials, those that have no analogue equivalent. And this risk is most clearly manifested by the Internet and whatever successors may emerge over time.

This is reflected in the Guidelines which note that 'this digital heritage is likely to become more important and more widespread over time' and 'new forms of expression and communication have emerged that did not exist previously. The Internet is one vast example of this phenomenon.'

I want now to note a few attributes of the web that together comprise a unique challenge to the digital preservation programme and which broaden the conversation regarding the impact of digital both globally and across all of our lives.

These attributes include:

¹⁰ Jesse Lichtenstein, "Why open data alone is not enough," *Wired Magazine* (July 2011), accessed August 24, 2012, http://www.wired.com/magazine/2011/06/st_essay_datafireworks/.

¹¹ Nancy Y. McGovern, ed., *Aligning national approaches to digital preservation* (Atlanta, GA: Educopia Institute Publications, 2012), accessed August 24 2012, http://www.educopia.org/sites/default/files/ANADP_Educopia_2012.pdf.

The potential to leave us with more and richer histories of moments in time - social history, minorities, ethnic communities - the nature and extent of heritage and history have the potential to be opened out.

The historian Henry Steel Commager stated about newspapers that “this is what really happened, reported by a free press to a free people. It is the raw material of history; it is the story of our own times.”¹² Increasingly this will be the role of the Internet.

The range of people engaging in digital culture far outstrips the range of people engaging in print culture. The nature of those participating in digital culture is substantially more diverse. More histories will be made available to us over time. These histories would not have left a trace in print culture.

The Guidelines note that ‘using computers and related tools, humans are creating and sharing digital resources - information, creative expression, ideas and knowledge ... that they value and want to share with others over time as well as across space.’ They also note that ‘increasingly this is a heritage that documents the actions of governments, the results of scientific research, the debate of ideas, the aspirations and imagination of communities, the histories of the current and coming world.’ Almost everyone, everywhere is online in one form or another, or will be.

Social media is about direct, instantaneous communication and can result in the creation of social movements which may never have been recorded in the past. Our digital collecting and our preservation programme need to reflect this.

Where recommendations used to be from semi-elitist sources – the food critic, the movie critic – the range of ‘experts’ that we rely on for input across the universe of our choices has increased dramatically. What are the implications for citizenship and an active democratic process? Do we face a democratisation of recommendation / information or are we heading towards a Tower of Babel? Will we become more engaged or will we become more passive?

The Guidelines note that ‘definitions of heritage need to be seen in context’, that ‘heritage value may also be based on what is important at a group or community level’, that ‘heritage materials can exist well beyond the limits suggested by national legislation or international conventions’ and, finally, that ‘anything that is considered important enough to be passed to the future can be considered to have heritage value of some kind.’

6. A Role for UNESCO?

In the above I have tried to show that the UNESCO Charter and Guidelines were exemplary documents of their time using the sections on Responsibilities and principles as a litmus. They traversed the core issues relating to the long-term safety of national digital heritage and provided a conceptual framework which is still valid today. I have also suggested, in the section on digital preservation that the promise of these

¹² New York Times, *One hundred years of famous pages from the New York Times, 1851-1951* (New York, NY: Simon and Schuster, 1951).

documents has not yet been met and that digital preservation practice still falls far short of the objectives that the authors of the Charter and Guidelines had envisaged.

In the following sections I want to show that UNESCO not only has a role to play in the digital preservation domain but that it has a unique role related to its global reach and objectives.

The potential to collect, preserve and make accessible a fuller expression of the cultures, heritages, histories of peoples is within our grasp. This is not to say that traditional activities related to selection and appraisal and the related assigning of value will not continue. However, it is to say that the opportunity exists in the digital sphere and within the parameters of a robust, scalable digital preservation programme to ensure that substantially more of the multiple histories of the world can be kept for the benefit of the future.

In an article for the *St Louis Post-Dispatch* Bill McClellan describes the agonising choice as to whether to keep a colleague and friend's letters and papers after he died and the serendipitous path that lead him to pass the papers on to the St. Louis Media Archives.¹³ This is not a particularly unusual occurrence in itself but how will this story be played out in the ubiquitous world of the internet? How will that ubiquity shape what history, what culture, what heritage of peoples' is still available and accessible to the future?

Why should UNESCO be a part of the global digital preservation programme? It is not an accident that national libraries, archives, museums are called memory institutions and the history of deliberate destruction of libraries is as long as the history of print culture itself.

- Destruction of the Library of Alexandria – date uncertain
- Destruction of scientific and philosophical library in Cordoba – 10th century
- Destruction of the Corvina Library in Buda – 1526
- Destruction of the Fatimid Library in Cairo – 1806
- Destruction of the libraries and archives of the Maya – Spanish Colonialists
- Destruction of the libraries and archives of the Aztecs – Spanish colonialists
- Destruction of the National and University Library of Bosnia and Herzegovina – 1992
- Destruction of the Abkhazian Research Institute of History, Language and Literature – 1992
- Mayor of Orange, France removing material deemed to be not truly French in support of far right National Front party – 1996
- Patriot Act, United States requires libraries to hand over details of their users – 2001
- Burning Harry Potter in New Mexico – 2001
- Destruction of National Library of Iraq including books that survived the sacking by the Mongols in 1258 when the waters of the Tigris were said to have run black with ink – 2003
- Court ordered burning of books in Cuba - 2005
- Dove World Outreach Centre, Florida – 2010, in the end refrains from burning the Koran.

What is the impulse behind these acts? What is it that is feared? If it is not the memory held within these institutions? This is why digital preservation matters and this is why UNESCO needs to be at the forefront of moves towards a global approach to digital preservation.

¹³ B. McClellan, "McClellan: Friend's letters now part of history," *St. Louis Post-Dispatch*, July 30 2012, accessed August 26, 2012, http://www.stltoday.com/news/local/columns/bill-mcclellan/mcclellan-friend-s-letters-now-part-of-history/article_b9c8158a-523c-575c-bc8a-67dcb95d131d.html.

7. A Goal for UNESCO?

Parts of a solution should include:

- We need to be more purposive about weaving digital preservation into the wider strategic approach to our digital activities;
- We need to engage more methodically with the increasing quantity and complexity of digital materials going forwards;
- We need secure, stable, agreed relationships with large institutional creators (e.g., newspaper publishers), academic and private research producers etc.;
- We need to move from short term project funding to ongoing sustainable funding recognising the ongoing-ness of digital preservation;
- We need to engage with the full spectrum of national and international stakeholders to make this work;
- We need to make explicit the long-tail implications of digital preservation.

From a UNESCO point of view the following quote from Marcus Garvey is characteristic of why digital preservation is important:

A people without the knowledge of their past ... is like a tree without roots.

From a national library point of view the following quote is characteristic of why digital preservation is important:

A National Library is a place where a nation nourishes its memory and exerts its imagination – where it connects with its past and invents its future.¹⁴

But even national libraries are not immune to the chill effects of government priorities, hostile economic requirements, and a lack of understanding that a nation's self-worth, identity, etc., are supported by national libraries.¹⁵ While this is one person's view of a very heated current debate it is a salutary reminder that offices of culture such as national libraries and national archives do not necessarily have a privileged position to protect them from the prevailing winds. The Guidelines state that 'making sure this burgeoning digital heritage remains available is thus a global issue relevant to all countries and communities.' In looking to refresh the excellent work of the 2003 documents, a larger goal for UNESCO could be: ***to see the vision of the Charter embedded in the national legislations of its member states.***

In this way UNESCO would be fulfilling its objectives for digital preservation as stated in Article 12 of the Charter and also reinforcing Article 19 of the Universal Declaration of Human Rights which states that 'everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.'¹⁶

¹⁴ Pierre Ryckmans, "Perplexities of an electronically illiterate old man," *Quadrant* 40, no. 9 (Sept 1996): 12-15.

¹⁵ V. Knowles, "Closing doors on Canada's history," *iPolitics*, August 10, 2012, accessed August 26, 2012, <http://www.ipolitics.ca/2012/08/10/val-knowles-closing-doors-on-canadas-history/>.

¹⁶ United Nations. *Universal Declaration of Human Rights*. *United Nations General Assembly, 10 December 1948*, accessed August 29, 2012, http://en.wikipedia.org/wiki/Universal_Declaration_of_Human_Rights.

The Economics of Preserving Digital Information

The Economics of Long-Term Digital Storage

David S. H. Rosenthal,¹ Daniel C. Rosenthal,² Ethan L. Miller,² Ian F. Adams,² Mark W. Storer³ and Erez Zadok⁴

¹Stanford University Libraries, USA, ²UC Santa Cruz, USA, ³NetApp, ⁴Stony Brook University, USA

Abstract

The cost per byte of storage media has dropped exponentially for three decades. Despite this, a Blue Ribbon panel described economic sustainability as the major issue facing long-term digital preservation. If economics are the major concern even when prices drop rapidly, what will happen if they slow or stop? We present evidence that they will, and some simulations of the impact on digital preservation costs.

Author

David Rosenthal is the Chief Scientist of the LOCKSS Program at the Stanford Library. Ethan Miller is Associate Director, and Daniel Rosenthal and Ian Adams are graduate students at U.C. Santa Cruz 's Storage Systems Research Center. Mark Storer is a Member of Technical Staff in the Advanced Technology Group at NetApp. Erez Zadok is Director of Stony Brook University's Filesystems and Storage Lab.

1. Introduction

Paper as the medium for the world's memory has one great advantage; it survives benign neglect well. Bits, on the other hand, need continual care, and thus a continual flow of money. A Blue Ribbon panel described economic sustainability as the major issue facing long-term digital preservation.¹ This is despite Kryder's Law, the 30-year history of the cost of digital storage media dropping exponentially. If economics are the major concern even when Kryder's Law holds, what will happen if it slows or stops?

We present the growing body of evidence suggesting that Kryder's Law will not be as helpful in the future as it has been in the past. The dimming prospect for Kryder's Law is a major motivation for research under way with participants from UC Santa Cruz's Storage Systems Research Center, Stony Brook University's Filesystems and Storage Lab, the LOCKSS² Program at the Stanford Libraries and NetApp. The goal is to develop a comprehensive economic model of long-term digital storage capable of being used for scenario planning by a wide range of digital archives, and which can be used as a component of broader models of digital preservation costs. Results from prototypes of this model illuminate issues important for preservation such as the impact of the recent disk price spike, and the cost-effectiveness of cloud storage.

2. Economics of Digital Preservation

There is a substantial body of work on the cost of digital preservation. Some does not, or not yet, cover storage costs:

¹ Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010.

² Lots Of Copies Keep Stuff Safe, a trademark of Stanford University.

- CMDP³ is an effort under way, funded by the Danish Ministry of Culture, to build a cost model for each of the activities identified in the OAIS reference model.⁴ Initial work focuses on the early activities, preservation planning and ingest; CMDP has yet to deal with long-term storage costs.
- The Blue Ribbon Task Force on Sustainable Digital Preservation and Access was funded by the NSF, the Andrew W. Mellon Foundation, the Library of Congress, the U.K. Joint Information Systems Committee (JISC), the National Archives and Records Administration, and the Council on Library and Information Resources. Their final report does not treat storage costs.⁵

Others include storage costs:

- LIFE is funded by the UK's JISC to build a life-cycle model of e-literature in a series of phases.⁶ Storage costs were first treated retrospectively in Phase 2⁷ and at a more detailed prospective level in Phase 3.⁸
- KRDS⁹ is also funded by JISC. It is primarily focused on identifying the value of digital collections, but in its initial phase¹⁰ it developed a cost model including storage costs.
- The PrestoPrime project funded by the EU developed an interactive simulation of preservation costs including storage costs.¹¹
- ENSURE is an EU-funded project in its early stages of building a preservation cost model "Based on cost data collection" that "aims to tackle the challenges that face cost modelling for long-term digital preservation."¹²
- Stephen Chapman¹³ compared historic storage costs for analogue items in the Harvard Depository with those for digital objects at OCLC.
- The California Digital Library has developed a Total Cost of Preservation model that includes storage costs.¹⁴

Storage costs are only one element of the total cost of digital preservation. These studies confirm that a significant part of the total, half in some studies, has to be paid up-front as content is ingested. But storage is important in each of these models as it is a large part of the continuing cost. The models these studies use to project future storage costs are based on collecting historical cost data and using it to project future costs. They implicitly assume that Kryder's Law continues in the future as it has in the past. If this assumption were not to hold it would have two significant effects:

³ Kejser, Nielsen and Thirifays, 2011.

⁴ ISO 14721:2003.

⁵ Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010.

⁶ Wheatley, Ayris, Davies, McLeod and Shenton 2007.

⁷ Ayris, Davies, McLeod, Miao, Shenton and Wheatley 2008.

⁸ Wheatley and Hole 2009.

⁹ Beagrie 2012.

¹⁰ Beagrie 2008.

¹¹ Addis and Jacyno 2010.

¹² Xue, Shehab, Baguley and Badawy 2011.

¹³ Chapman 2006.

¹⁴ Abrams, Cruse, Kunze and Mandrane 2012.

- The proportion of overall of digital preservation costs represented by storage costs would greatly increase, since the cost of storing any individual object would no longer rapidly become insignificant.
- The projected total future cost of digital preservation would rise significantly.

If Kryder's law will not continue the current cost forecasting techniques will produce misleadingly optimistic projections, leading to increased risk of economic failure. We need a new approach to modeling the storage cost component of the overall cost of digital preservation.

Strictly, Kryder's Law is not about cost. It states that the areal density of bits on disk platters roughly doubles every two years.¹⁵ The cost implication of this was popularized by Clayton Christensen's 1997 book *The Innovator's Dilemma*¹⁶ but it has actually held for about three decades. Until very recently, the disk drive business was highly competitive, with no manufacturer having a dominant market share, so increases in areal density resulted in corresponding decreases in cost per bit. In practice, consumers got double the capacity at approximately the same price every two years.

3. Storage Technology Futures

Unfortunately, there is a growing body of evidence suggesting that future improvements in storage cost per bit will be much slower than in the past. This applies to disk, tape and the various forms of solid state storage. IDC's projections for the storage industry as a whole¹⁷ show slowing in both the rate of decrease in cost per bit and in the rate of investment in digital storage through 2015.

3.1 Disk

In 2011 disk represented 70% of all the bytes of storage shipped.¹⁸ The disk industry's roadmap used to predict a consistent 40%/yr improvement in bit density on disk platters, which translated to a 40%/yr reduction in cost per bit stored.¹⁹ In recent years the industry has failed to achieve this roadmap target.²⁰ The current roadmap predicts no more than a 20%/yr improvement in bit density for the next five years.²¹ There are reasons to believe that even this may be optimistic, and also that even if it were to be achieved it might not translate directly into a 20%/yr drop in cost per bit:

- The disk drive industry used to be a commodity business marked by intense competition and low margins. Eventually, the weaker players became vulnerable and in 2012 a spate of mergers transformed the industry.²² Western Digital and Seagate now have more than 85% of the market,²³ so there is much less competition. The market is expected to support considerably

¹⁵ Walter 2005.

¹⁶ Christensen 1997.

¹⁷ Gantz, Manfrediz, Minton, Reinsel, Schlichting and Toncheva 2011.

¹⁸ Mellor 2012.

¹⁹ Walter 2005.

²⁰ Mellor 2010.

²¹ IHS iSuppli 2012.

²² Brandom 2012.

²³ IHS iSuppli 2012.

higher margins in the future. An increase in margins represents an effective reduction in the future rate of cost drop.

- The recording technology used by the most recent five generations of disk drives is Perpendicular Magnetic Recording (PMR). According to earlier versions of the industry roadmap, it should have been replaced by Heat Assisted Magnetic Recording (HAMR) by early 2010. HAMR uses a laser to heat the magnetic material on the platter to reduce the size of the area whose magnetism is changed by a write operation. The transition from PMR to HAMR has been delayed because it has turned out to be vastly more difficult and expensive than was predicted. The cost of this transition was a factor driving the consolidation of the industry.
- Unable to deploy HAMR, the industry has resorted to what can only be described as desperate measures to stretch PMR into a sixth generation using a technique called *shingled writes*. This involves writing tracks on the disk so close together that they overlap, and using sophisticated signal processing techniques to disentangle them on a read. This causes system-level problems because disks are no longer randomly writable, they become in effect append-only devices.²⁴ Mitigating these problems by adding capabilities to the disk hardware increases cost and reduces capacity; addressing them by changing operating systems is expensive and disruptive. Shingled write technology may not be a way for disks to stay on the Kryder's Law curve for another technology generation.
- In the past the disk industry has responded to difficulty in increasing bit density and thus in offering higher capacity in the same form factor by adding platters.²⁵ Since adding platters adds cost, and very few more platters can be added without disruptive and expensive changes in the drive form factor, this is less effective than increasing bit density in decreasing cost per bit.
- The favored successor technology to HAMR is Bit-Patterned Media (BPM), which uses lithographic techniques to create an extremely small location for each bit on the platter. The transition from HAMR to BPM is now expected to be even more difficult and expensive than the PMR to HAMR transition. It is therefore likely to encounter similar delays, which act to reduce the rate of cost drop.
- As magnetic particles on the platter get smaller, the temperature below which they can retain information for a given time decreases.²⁶ 'The miniaturization of magnetic recording devices, which store information in nanosized magnetic grains or "bits" is constrained by the so-called superparamagnetic limit: when grains are too small, thermal fluctuations can easily flip the direction of magnetization in each bit, causing permanent loss of information.' For the temperatures and times involved in disk storage, this is expected to limit bit densities to well under 100Tb per square inch.²⁷ At the current roadmap's 20%/yr density increase, this limit could be encountered as soon as 2030; at the 40%/yr used by e.g., Kryder and Kim (2009)²⁸ it would be encountered sometime after 2022. It is to be expected that the rate of increase in bit density will slow as the limit is approached; the current slowing may be early evidence of this.

²⁴ Amer, Holliday, Long, Miller, Paris and Schwarz, 2011.

²⁵ Mellor, 2010.

²⁶ Melikyan, 2008.

²⁷ Shiroishi, Fukuda, Tagawa, Iwasaki, Takenoiri, Tanaka, Mutoh and Yoshikawa, 2009.

²⁸ Kryder and Kim, 2009.

- Most disks used for storing long-term data are consumer 3.5" SATA drives, providing large capacity per drive with reasonable performance and good reliability.²⁹ Because they were an essential component of consumer desktop PCs, they have had very large manufacturing volumes and thus very low costs. The consumer PC market has moved to laptops, which use 2.5" drives, and is moving to tablets and ultrabooks, which use flash memory. A total of 415M PCs sold in 2011, of which 66M were tablets (growing at 274%/yr) and about 210M were laptops (growing at 7%) including netbooks and ultrabooks. That leaves 139M desktops (growing at 2%) and servers, which are candidates for 3.5" drives.³⁰ 2.5" disks use the same recording technology as 3.5" disks, and their cost per bit has been decreasing at a similar rate, but they are typically 3-4 years later than 3.5" disks at reaching a particular \$/GB value³¹ and thus, at the historic 40%/yr price drop, 3-5 times as expensive. 3.5" drives consequent loss of manufacturing volume will probably slow the cost drop for long-term data. If long-term data migrates to 2.5" drives it will suffer a significant cost increase. Because both form factors are on parallel Kryder's Law curves, 2.5" drives will never catch up with where 3.5" drives would be if they still existed. By the time this migration occurs, the consumer laptop market for 2.5" drives will probably be in eclipse, reducing their manufacturing volumes too.
- Disk industry insiders³² regard HAMR as much more suitable for 2.5" than for 3.5" drives. If it is initially deployed only on 2.5" drives, this will drive long-term data from 3.5" to 2.5" drives more quickly, making the price increase sharper.
- The 2011 floods in Thailand destroyed about 40% of the world's disk drive manufacturing capacity. Disk drive prices doubled almost overnight, and have yet to return to pre-flood levels,³³ let alone to the levels to which they would have dropped absent the floods. Part of the reason is the enormous cost to the industry of replacing the lost capacity,³⁴ but an additional reason is that the disk manufacturing duopoly has seized the opportunity to increase their margins, which were about 6% for Western Digital and 3% for Seagate pre-flood and are now about 16% and 37%, respectively.³⁵

3.2 Tape

Tape is an important medium for long-stem storage of large amounts of data. At scale, i.e., in large tape robots, its low media costs, low power consumption and relatively high reliability outweigh its long access times. The recording technology used by tape lags about 8 years behind disk, but it is on approximately the same cost per bit curve as disk.

However, tape's share of the total storage market is shrinking, which means it will get less of the total storage R&D investment pool than it used to. Thus we can expect tape's cost per bit to continue dropping, albeit somewhat more slowly than previously, for perhaps another 8 years. This will

²⁹ Pinheiro, Weber and Barroso, 2007; Schroeder and Gibson, 2007.

³⁰ McKendrick, 2012, delboy, 2012.

³¹ Grochowski and Fontana, 2011.

³² Anderson, 2009.

³³ Kunert, May 2012.

³⁴ Kunert, July 2012.

³⁵ Hruska, 2012.

significantly increase tape's cost advantage over disk while it happens, although the technological issues of disk recording technology will eventually affect tape too.

3.3 Solid State Memories

Flash memory is currently much more expensive per byte than hard disk but because of its other attributes, low power, small form factor, robustness, it has captured a significant part of the storage market. It may be that these attributes, which are also important for long-term storage,³⁶ will drive flash into that market too. However, there are a large number of alternative solid state technologies on the horizon, some of which are even more promising for long-term storage than flash. Kryder and Kim (2009)³⁷ surveyed the prospects in 2020 for both flash and the alternative solid state technologies, comparing them with their projection for hard disk technology at that date based on a 40% annual increase in bit density. At this rate in 2020 hard disk would still have a factor of 3-10 in bit density to go before reaching the superparamagnetic limit. They conclude:

...to compete with hard drives on a cost per terabyte basis will be challenging for any solid state technology, because the ITRS lithography roadmap limits the density that most alternative technologies can achieve.

Adjusting their projections for a 20% annual increase in hard drive bit density reduces the 2020 target from 10Tb/in² to 1.8Tb/in², or from a 40TB to a 7TB 2.5" drive. This would probably lead to solid state technology capturing more of the storage market, and thus more of the R&D investment, than Kryder and Kim assume, reducing still further hard disk's competitiveness. It would not, however, change their basic conclusion that competing with hard disk on a cost per bit basis would be a challenge. By 2020 all the solid state technologies they surveyed would be approaching technological limits, whereas the lower bit density growth rate implies that then hard disks would still be a factor of 15-50, or 1-2 decades from the superparamagnetic limit.

Thus, although we may expect solid state technology to become more cost-competitive with hard disk in the short term, by the end of this decade this competitiveness will probably decline.

4. Storage Business Models

There are three fundamentally different business models for long-term storage. None of them has the properties a customer would like:

- It can be *rented*. For example, Amazon's Simple Storage Service (S3) charges \$0.125 per GB per month with discounts for large volumes.³⁸ This rent can be decreased or even increased over time, so from the service's point of view the model is not dependent on the Kryder's Law decrease. From the customer's point of view, this model is risky. Unexpected rent increases or even temporary fluctuations in the customer's money supply can lead to permanent loss of data due to inability to pay the monthly rent. Each access to data in S3 costs on the order of a month's

³⁶ Adams, Miller and Rosenthal, 2011.

³⁷ Kryder and Kim, 2009.

³⁸ Amazon, 2012.

storage; customers could be in the awkward position of being able to pay for their data to be stored but being unable to afford to access it.

- The stored content can be *monetized*. For example, Gmail offers a gradually increasing amount of e-mail storage free to users. Google makes money by selling ads when the user accesses their mail. As each message gets older, it is accessed less and less frequently, as is common in archived data. Thus Google makes less and less money from older and older mail, meaning that the Kryder's Law decrease in the cost per bit of storing old mail is important to this business model. But it is not essential. Google can adjust the rate at which it supplies storage to users, reduce their storage allocation, or even start charging users who never click on ads for their e-mail storage, to match their cost of storage and the income from advertisements over time. The customer has no leverage over the service, making it risky for them. The survival of the data is at the whim of the service; if it no longer makes money from the data it will no longer be motivated to preserve it.
- The stored content can be *endowed*, deposited in the storage service together with a sum of money thought to be sufficient to pay for its storage through its entire life. Determining an appropriate sum involves projecting both the Kryder's Law decrease in cost and the future interest rate which will apply to the unexpended part of the endowment. If these projections turn out to be too low the data is at risk, since the service will not be able to afford to keep it.

5. Discounted Cash Flow

For the purpose of building models the endowment approach has a great advantage. It provides an apples-to-apples way to compare the flows of money through time. In effect, it uses the economists' standard technique for doing so, Discounted Cash Flow (DCF). DCF computes the *net present value* of a future expenditure by assuming a constant interest rate, the *discount rate*, and computing the amount less than the future expenditure which, with the addition of the interest accumulated by then, would amount to the future expenditure when it occurs.

Recent research has thrown serious doubt upon both the practical usefulness and theoretical basis of DCF. Its practical usefulness is suspect because it involves choosing a discount rate that will apply for the duration. In practice, people applying DCF choose unrealistically high interest rates, making investment in long-term projects much more difficult to justify than it should be.³⁹ Its theoretical basis is suspect because the single constant interest rate averages out the effect of periods of very high or (as now) very low interest rates. This would be correct if the outcome was linearly related to the interest rate, but it isn't.⁴⁰ This non-linear behavior implies that Monte Carlo models are required to compute the net present value of expenditures.

6. Economic Models of Storage

The difference between the net present value computed by DCF and that computed by models including variable interest rates increases through time, making DCF less and less useful for analysing storage costs the longer the duration of storage.

³⁹ Haldane and Davies, 2011.

⁴⁰ Farmer and Geanakoplos, 2009.

The inadequacy of DCF and the prospect of no longer being able to count on the rapid decrease in cost per bit to make long-term storage costs insignificant motivated our work to develop a Monte Carlo model. The goal is to understand the impact of changes in Kryder's Law and other factors such as interest rates on the cost of storing data.

This work is at an early stage. To explore the problem space, and to communicate with potential users, we have developed some prototype models. The prototypes are producing plausible results, but they have not been validated against real-world data, so their results should not be relied upon. With experience from the prototypes and the feedback we have received from potential users we are developing an integrated, comprehensive model. In the meantime we present results from two of the prototypes.

6.1 Short-Term Model

Our first model follows a unit of hardware, as it might be a shelf of a filer in a data center, facing an exponentially growing demand to store data. Disks are added as needed; their capacities grow over time according to Kryder's Law. They consume power and labor, and are replaced as they fail or end their service life.

Figure 1 shows an example analysis. Parameters are set to plausible values, such as 57% annual growth in demand for data storage⁴¹ and 5% probability of failure in service (estimated from Pinheiro et al.⁴²). The graph shows how total cost of ownership and its components vary with the service life of the disks. It demonstrates the well-known observation that disks (and tapes) are replaced when their density becomes too low to justify the space and power they use, not when their life expires. With our chosen parameters, the model predicts optimum disk replacement in under 3 years.

Note the counter-intuitive increase of the labor component with increasing service life of the disks. Each disk is assumed to consume a fixed amount of labor to install and replace. As service life increases the total number of drives in use increases; the demand for storage remains the same but the average drive becomes smaller and older.

6.2 Long-Term Model

The long-term prototype follows a unit of data through time as it migrates from one storage medium to another, taking up a smaller and smaller proportion of each successive medium. It computes the *endowment* needed to preserve the data using a model of interest rates based on the 20-year history of inflation-protected US Treasury bonds.⁴³

This model includes storage media, with purchase and running costs. They are replaced with successor media when their service life expires, or when new media become available whose costs are enough lower to justify the cost of migrating out of the old medium into the newer one. The endowment earns interest, and pays for the purchase, running and migration costs.

Different models of interest rate variation through time can greatly affect the endowment computation,⁴⁴ so the absolute value of the endowment computed by the long-term prototype should be treated skeptically. Nevertheless, with similar parameters this model produces similar endowment values

⁴¹ Gantz, Manfrediz, Minton, Reinsel, Schlichting and Toncheva, 2011.

⁴² Pinheiro, Weber and Barroso, 2007.

⁴³ US Dept. of Treasury, 2012.

⁴⁴ Farmer and Geanakoplos, 2009.

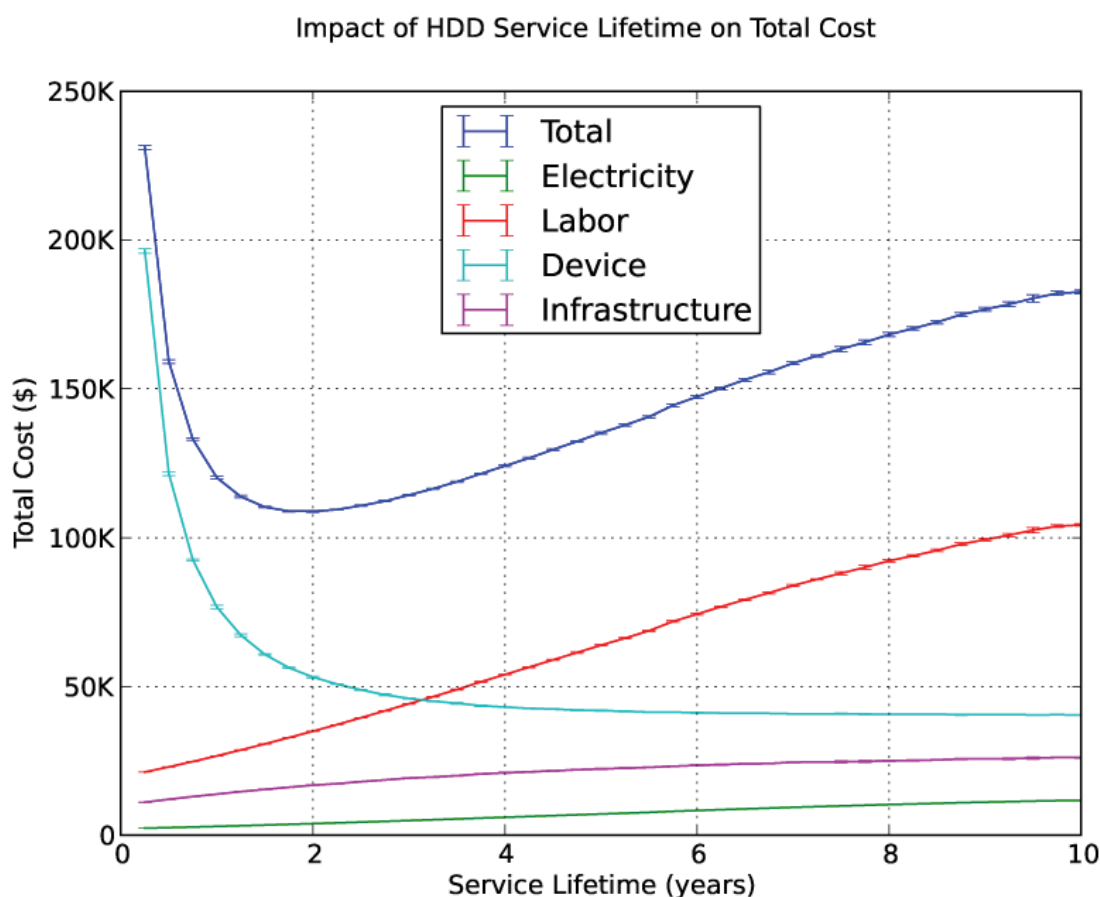


Figure 1. The effect of hard drive service lifetime on 10-year cost of ownership of an archive growing 57%/yr.

to those of other approaches, e.g., Abrams (2012)⁴⁵ and Addis (2010).⁴⁶ When comparing the effect of other parameters through time, for example different storage technologies or media replacement policies, we use the same interest rate model for each, so their relative cost is unaffected.

6.3 Varying Kryder's Law Rates

Figure 2 shows a simple output of this model, plotting the probability that the data will last 100 years without running out of money (Y axis) against the endowment as a multiple of the initial storage cost for a fixed Kryder's Law rate, in this case 25%/yr. Interest rates are modeled on the past 20 years, and the service life of the media is 4 years. As one would expect, it is an S-curve. If the endowment is too small, running out of money is certain. If it is large enough, survival is certain. The insight from this graph is that the transition from 0% to 100% survival takes place over only about 10% of the critical endowment value. A 25%/yr Kryder's Law rate dominates the effect of the much lower interest rates.

⁴⁵ Abrams, Cruse, Kunze and Mandrane, 2012.

⁴⁶ Addis and Jacyno, 2010.

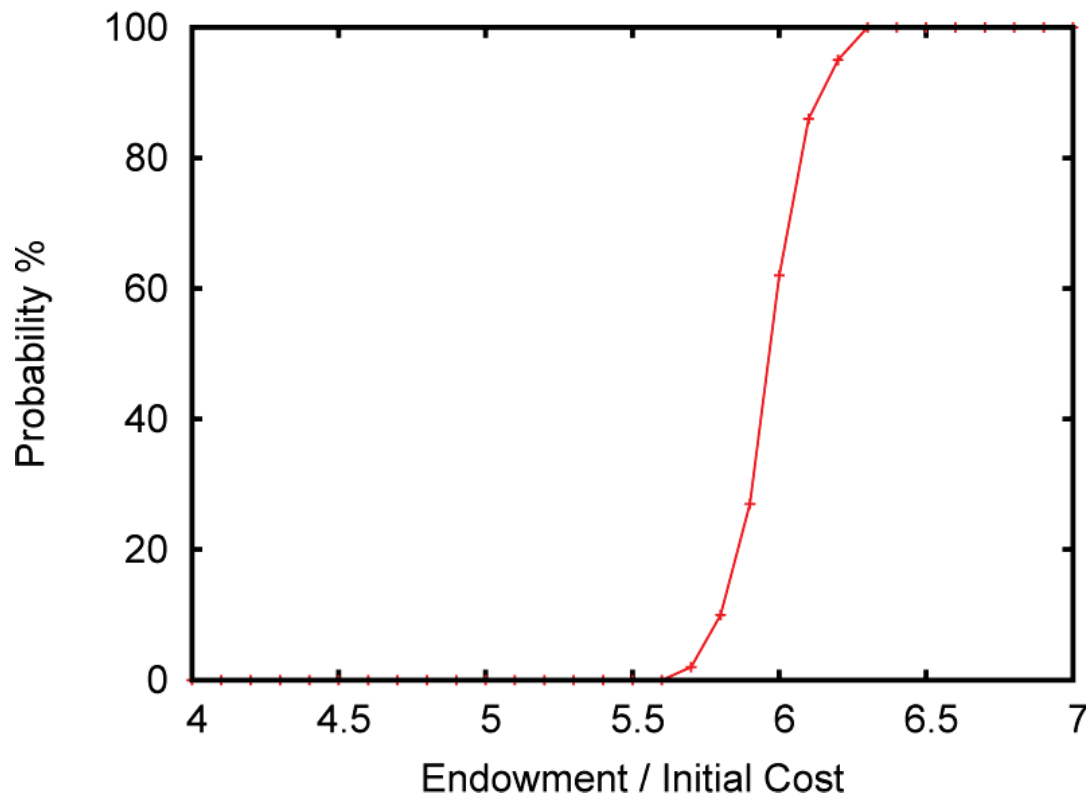


Figure 2. Survival Probability vs. Endowment as a multiple of initial cost.

Repeating the simulation for a range of Kryder's Law rates gives the surface shown as a heat map in Figure 3. Note that the transition is less abrupt the lower the Kryder rate.

Taking the 98% survival probability contour of this surface gives the graph of Figure 4. One insight from this graph is that historic Kryder's Law rates have been on the flatter, right hand side of the graph, where their effect on the endowment needed is small. The values we expect in the future are on the steep, left hand side of the graph, where the endowment needed is much larger and depends sensitively on the Kryder's Law value. Thus we will be moving from an era when storage was affordable and predicting future storage costs was less important to an era when storage is expensive and predicting future costs is very important.

6.4 Price Spikes

Figure 5 shows an example analysis of the impact of a spike in disk costs such as that caused by the recent floods in Thailand. Interest rates are modeled on the past 20 years, media costs drop exponentially at various rates, and the service life of the media is 4 years. After a variable delay, media costs double for a year then resume their exponential decrease. The graph shows the endowment that provides 95% probability of surviving 100 years without exhausting it, as a multiple of the initial cost. Plausibly, if storage costs drop rapidly spikes have little effect but if they drop slowly the effect is large. Also, if costs drop slowly enough that media are replaced at their service life, and the spike happens at that time, the effect is amplified. We can expect global warming to increase the frequency of supply chain disruptions, and Kryder's Law to flatten, so the impact of future price spikes will be probably be significant.

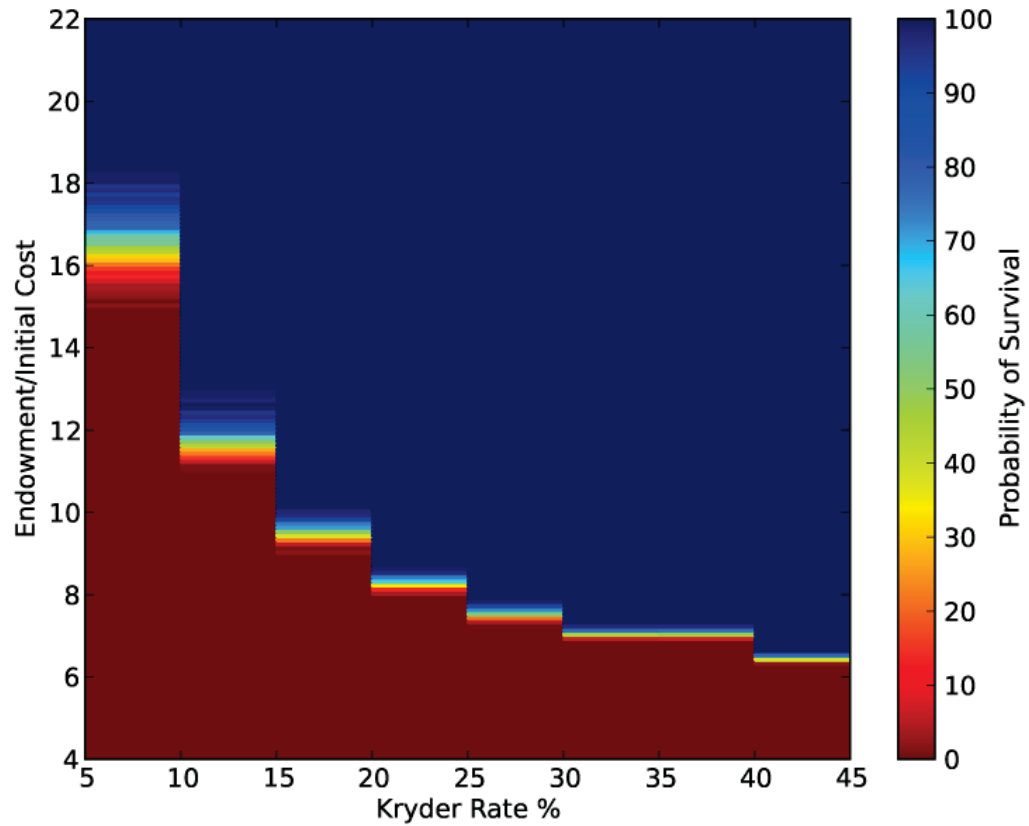


Figure 3. Survival Probability vs. Endowment vs. Kryder rate.

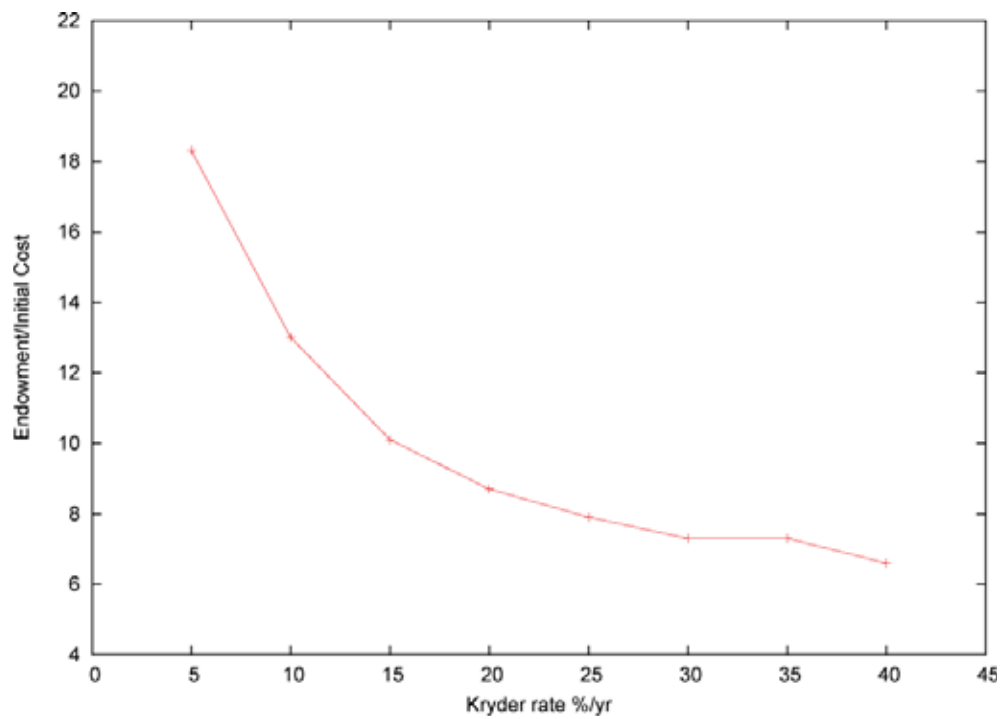


Figure 4. Endowment vs. Kryder rate.

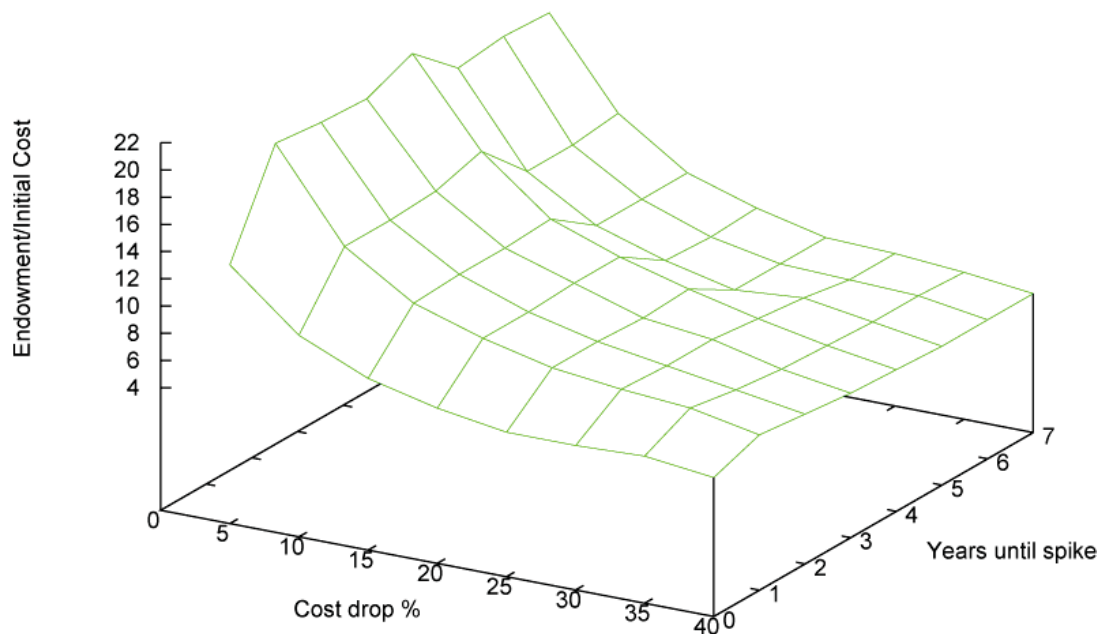


Figure 5. The impact of spikes in media cost on the endowment required for 95% survival at various Kryder rates. Zero delay has no spike for comparison.

6.5 Solid State Storage

As we see from the short-term model, the raw media cost is only a part of the total cost of storage, even with a relatively short 10-year time horizon. The cost differential between flash and hard disk has been decreasing as flash gains market share. For long-term storage flash has advantages including low power consumption, small form factor, physical robustness and long device lifetime. Suitably exploited,⁴⁷ these factors can outweigh the higher purchase price and deliver lower total cost of ownership over a period.

In principle, at times of low interest rates (such as now) it makes sense to invest in storage technologies with higher capital cost but lower running costs and long lifetimes. At times of high interest rates, it makes sense to invest in technologies with lower capital costs and higher running costs and short lifetimes.

Unfortunately, for an organization to justify investing in solid state storage on this basis requires that it have both a long enough planning horizon and an accounting policy that distinguishes between capital and operating costs. Many organizations lack both; for example most University libraries run on annual budget cycles, are not allowed to carry reserves from year to year, and cannot borrow to finance equipment purchases. Thus, even if solid state storage could offer lower total cost of ownership over say 5 years, they would be unable to invest to capture these savings. This is an example of the problem of short-termism identified by Haldane and Davies (2011).⁴⁸

Our long-term model includes a planning horizon parameter, but we have not yet been able to conduct a detailed study of its effect on investing in solid state storage.

⁴⁷ Adams, Miller and Rosenthal, 2011.

⁴⁸ Haldane and Davies, 2011.

6.6 Cloud Storage

Table 1 shows the history of the prices charged by several major storage services. It shows that they drop at most 3%/yr. This is in stark contrast with the 30-year history of raw disk prices, which have dropped at least 30%/yr.

Table 1. Price History of Some Storage Services

Service	Launch date	Launch \$/GB/mo	2012 \$/GB/mo	Price Drop %/yr
Amazon S3	Mar '06	0.15	0.13	3
Rackspace	May '08	0.15	0.15	0
Azure	Nov '09	0.15	0.14	3
Google	Oct '11	0.13	0.13	0

This comparison is somewhat unfair to S3. Amazon has used the decrease in storage costs to implement a tiered pricing model; over time larger and larger tiers with lower prices have been introduced. The price of the largest tier, now 5PB, has dropped about 10% per year; prices of each tier once introduced have been stable or dropped slowly.

Nevertheless, it is clear that the benefits of the decrease in raw storage prices are not going to cloud storage customers. Backblaze provides unlimited backup for personal computers for a fixed price, currently \$5/mo. Before the floods in Thailand, they documented the build cost of their custom storage hardware at under \$8K for 135TB.⁴⁹ They claimed their 3-year cost of ownership of a Petabyte was under \$100K; even S3's lower-cost Reduced Redundancy Storage (RRS) would have charged \$2.3M over the same period. Adjusting for the current 60% increase in disk prices since the floods⁵⁰ would make the build cost \$11.2K. Given S3's dominance of the cloud storage market, and thus purchasing volumes, it is very unlikely that their costs are much higher than Backblaze's. Despite this, 135TB in S3-RRS costs more than \$10K/mo. In the first month, an S3-RRS customer would pay almost as much as it would currently cost to buy the necessary hardware.

Why is cloud storage so expensive? Actually, in many cases, it isn't. Many customers have data whose life is much less than the life of the hardware, so they cannot amortize a hardware purchase over its life. Many customers, for example startup companies, have a very high cost of capital. Amazon and its competitors price against the value they deliver to these customers; not against their costs.

But Amazon and its competitors should be riding the Kryder's Law curve like everyone else. Why aren't they reducing their prices? Because they don't have to. Suppose you have 135TB in S3-RRS and you decide you are paying too much. You need to move your data somewhere cheaper. You are going to take a month to do it. It will cost you \$10,750, more than a month's storage, in bandwidth charges to get your data out,⁵¹ let alone the staff and other costs of doing the transfer, and checking that it worked

⁴⁹ Nufire, 2011.

⁵⁰ Kunert, May 2012.

⁵¹ Amazon, 2012.

correctly. A competitor is going to have to be a great deal cheaper than S3 to motivate you to pay these transition costs. Since S3 has the vast majority of the market, their costs are probably lower than any competitor's. If a competitor cut prices enough to take significant market share from S3, Amazon would undercut them.

7. Conclusions

From the foregoing, we can draw the following conclusions:

1. Optimistically, for the rest of this decade the rapid decrease in cost per bit of storage that has been a constant of the last three decades will be much slower; it might even stop.
2. This will make the expenditure commitment implied by a decision to preserve some digital content (a) much bigger and (b) much harder to predict than would be expected on the basis of history.
3. In a period of economic stringency, this increases the importance of developing accurate, predictive models of storage and other preservation costs.
4. For much of this decade tape is likely to maintain or improve its existing cost advantage over disk.
5. If organizations can change their accounting methods to properly recognize the long-term cost of ownership of preserved data, current low interest rates provide an opportunity to invest in solid state technologies which, despite their higher capital cost, are for this decade likely to provide lower total cost than disk, while retaining its rapid access.
6. The pricing models of current commercial cloud storage services are not suitable for long-term storage.

Acknowledgments

David Rosenthal worked under contract GA10C0061 from the Library of Congress' NDIIPP program. This research was supported in part by the National Science Foundation under awards CNS-0917396 (part of the American Recovery and Reinvestment Act of 2009 [Public Law 111-5]) and IIP-0934401, and by the industrial sponsors of the UC Santa Cruz SSRC, including EMC, HP, Hitachi, Huawei, IBM, LSI, NetApp, Northrop Grumman, Permabit, and Samsung.

Special thanks are due to Leslie Johnston and Jane Mandelbaum of the Library of Congress, and to Tom Lipkis, Thib Guicherd-Callin and Daniel Vargas of the LOCKSS Program.

References

- Abrams, S., P. Cruse, J. Kunze, and M. Mundrane. "Total Cost of Preservation (TCP): Cost Modeling for Sustainable Services." April 2012. <https://wiki.ucop.edu/download/attachments/163610649/TCP-total-cost-of-preservation.pdf>.

- Adams, I., E. L. Miller, and D. S. H. Rosenthal. "Using Storage Class Memory for Archives with DAWN, a Durable Array of Wimpy Nodes." Technical Report UCSC-SSRC-11-07, University of California, Santa Cruz, October 2011.
- Addis, M., and M. Jacyno. "Tools for modelling and simulating migration-based preservation." December 2010.
https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.2_PreservationModellingTools_R0_v1.00.pdf
- Amazon. "Simple Storage Service." 2012. <http://aws.amazon.com/s3/#pricing>.
- Amer, A., J. Holliday, D. D. E. Long, E. L. Miller, J.-F. Pâris, and T. Schwarz. "Data management and layout for shingled magnetic recording." *IEEE Transactions on Magnetics* 47, no. 10 (2011).
- Anderson, D. "Hard Drive Directions." September 2009.
http://www.digitalpreservation.gov/news/events/other_meetings/storage09/docs/2-4_Anderson-seagate-v3_HDtrends.pdf.
- Ayris, P., R. Davies, R. McLeod, R. Miao, H. Shenton, and P. Wheatley. "The LIFE2 Final Project Report." 2008. <http://eprints.ucl.ac.uk/11758/>.
- Beagrie, N. "Keeping Research Data Safe (Phase 1)." May 2008.
<http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx>.
- Beagrie, N. "KRDS/I2S2 Digital Preservation Benefit Analysis Tools Project." 2012.
<http://beagrie.com/klds-i2s2.php>.
- Blue Ribbon Task Force on Sustainable Digital Preservation and Access. "Sustainable Economics for a Digital Planet." April 2010. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.
- Brandom, R. "How Three Hard Drive Companies Gobbled Up The Industry." June 2010.
<http://www.buzzfeed.com/tommywilhelm/how-three-hard-drive-companies-gobbled-up-the-indu>.
- CCDS. *ISO 14721:2003: Reference model for an open archival information system*. 2003.
http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683.
- Chapman, S. "Counting the costs of digital preservation: Is repository storage affordable?" *Journal of Digital Information* 4, no. 2 (2006).
- Christensen, C. M. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business School Press, 1997.
- delboy. "66 Million Tablets Sold In 2011 Worldwide Led By iPad." 2012.
<http://www.worldtvpc.com/blog/67-million-tablets-sold-2011-world/>.
- Farmer, J. D., and J. Geanakoplos. "Hyperbolic Discounting is Rational: Valuing the Far Future With Uncertain Discount Rates." Technical Report 1719, Cowles Foundation, Yale University, 2009.
- Gantz, J. F., A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, and A. Toncheva. "The diverse and exploding digital universe." March 2011. <http://www.emc.com/collateral/about/news/idc-emc-digital-universe-2011-infographic.pdf>.
- Grochowski, E., and R. E. Fontana. "An Analysis of Flash and HDD Technology Trends." August 2011.
http://www.flashmemorysummit.com/English/Collaterals/Proceedings/2011/20110809_F2C_GrochowskiFontana.pdf.
- Haldane, A. G., and R. Davies. "The Short Long." In *29th Societe Univer sitaire Europeene de Recherches Financieres Colloquium: New Paradigms in Money and Finance?*, May 2011.

- Hruska, J. "HDD Pricewatch: Higher prices are the new normal." May 2012.
<http://www.extremetech.com/computing/129874-hdd-pricewatch-higher-prices-are-the-new-normal>
- IHS iSuppli. "HDD Areal Density Doubling in Five Years." May 2012.
<http://www.storagenewsletter.com/news/marketreport/ihs-isuppli-storage-space>.
- McKendrick, J. "Milestone: more smartphones than PCs sold in 2011." February 2012.
<http://www.smartplanet.com/blog/business-brains/milestone-more-smartphones-than-pcs-sold-in-2011/21828>
- Kejser, U. B., A. B. Nielsen, and A. Thirifays. "Cost Model for Digital Preservation: Cost of Digital Migration." *International Journal of Digital Curation* 6, no. 1 (2011).
- Kryder, M. H., and C. S. Kim. "After Hard Drives - What Comes Next?" *IEEE Trans. on Magnetics* 45, no. 10 (October 2009).
- Kunert, P. "Hard disk drive prices quick to rise, slow to fall." *The Register* (May 2012).
- Kunert, P. "WD: HDD prices won't fall to pre-flood levels until 2013." *The Register* (July 2012).
- Melikyan, A. "Synopsis: Pushing the superparamagnetic limit." 2008. <http://physics.aps.org/synopsis-for/10.1103/PhysRevB.78.214424>.
- Mellor, C. "Drive suppliers hit capacity increase difficulties," *The Register* (July 2010).
- Mellor, C. "WD bigshots spin superfast disk roadmap," *The Register* (May 2012).
- Nufire, T. "Petabytes on a Budget v2.0: Revealing More Secrets." July 2011.
<http://blog.backblaze.com/2011/07/20/petabytes-on-a-budget-v2-0revealing-more-secrets/>.
- Pinheiro, E., W.-D. Weber, and L. A. Barroso. "Failure Trends in a Large Disk Drive Population." In *Proceedings of 5th USENIX Conf. on File and Storage Technologies, San Jose, CA, 2007*.
- Schroeder, B., and G. Gibson. "Disk failures in the real world: What Does an MTTF of 1,000,000 Hours Mean to You?" In *Proceedings of 5th USENIX Conf. on File and Storage Technologies, San Jose, CA, 2007*.
- Shiroishi, Y., K. Fukuda, I. Tagawa, H. Iwasaki, S. Takenoiri, H. Tanaka, H. Mutoh, and N. Yoshikawa. "Future options for hdd storage." *IEEE Transactions on Magnetics* 45, no. 10 (2009).
- US Dept. of Treasury. "Daily Treasury Yield Curve Rates." 2012. <http://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/default.aspx>.
- Walter, C. "Kryder's Law," *Scientific American* (August 2005): 293.
- Wheatley, P., P. Ayriss, R. Davies, R. McLeod, and H. Shenton. "LIFE: costing the digital preservation lifecycle." In *iPRES2007*, 2007.
- Wheatley, P., and Hole, B. "LIFE3: Predicting Long-term Digital Preservation Costs." In *iPRES2009*, 2009.
- Xue, P., E. Shehab, P. Baguley, and M. Badawy. "Cost modelling for long-term digital preservation: Challenges and issues." In *Proc. 9th Intl. Conf. on Manufacturing Research ICMR2011*, 187–192, 2011.

Modelling the Costs of Preserving Digital Assets

Ulla Bøgvad Kejser,¹ Anders Bo Nielsen and Alex Thirifays²

¹The Royal Library, Copenhagen, Denmark, ubk@kb.dk

²The Danish National Archives, Copenhagen, Denmark, ubk@kb.dk and ubk@kb.dk

Abstract

Information is increasingly being produced in digital form, and some of it must be preserved for the long-term. Digital preservation includes a series of actively managed activities that require on-going funding. To obtain sufficient resources, there is a need for assessing the costs and the benefits accrued by preserving the assets. Cost data is also needed for optimizing activities and comparing the costs of different preservation alternatives. The purpose of this study is to analyse generic requirements for modelling the cost of preserving digital assets. The analysis was based on experiences from a Danish project to develop a cost model. It was found that a generic cost model should account for the nature of the organisation and the assets to be preserved, and for all major preservation activities and cost drivers. In addition, it should describe accounting principles. It was proposed to develop a generic cost model with specific profiles to represent different preservation scenarios.

Authors

Ulla Bøgvad Kejser holds a Ph.D. in Conservation-Restoration Science from the Royal Danish Academy of Fine Arts, School of Conservation. She is a preservation specialist at the Royal Library in Denmark working in the areas of preservation imaging, preservation planning, and curation of both physical and digital collections, especially images. Her current research focuses on modelling the costs of digital preservation.

Anders Bo Nielsen holds a master's in economics from the University of Copenhagen (economic history and IT as specialty) from 1997 and a master in IT (software development as specialty) from the IT University of Copenhagen from 2007. He is a principal consultant at the Department of Appraisal and Transfer at the Danish National Archives, where he has been employed since 1997.

Alex Thirifays is an archivist at The Danish National Archives. He obtained his MA in history in 1999. He has been working as an IT consultant for 6 years, and has, since 2007, been working with digital curation, particularly cost modelling, preservation planning and strategic issues. He developed the Cost Model for Digital Preservation (CMDP) together with Anders Bo Nielsen and Ulla Bøgvad Kejser and published several papers and reports on this work.

1. Introduction

All over the world societies produce more and more information in digital form, either directly from computers and other electronic devices or by digitization of analogue assets. Some of this information is considered worth preserving, e.g., critical business information, personal memories or cultural heritage documents. Preservation and access to digital assets is dependent on technology; the physical streams of bits constituting the information must remain intact and accessible through current systems, and the intellectual content of the information must remain usable over time. Counteracting the constant threat from technological obsolescence requires a series of actively managed activities, including emulation of systems or migration of formats. Facilities with adequate hardware and software must be in place as well as power to keep the systems running. In addition, people with the necessary skills to maintain the

systems and act as curators of the digital assets must be available. The required preservation activities and the flow of information are described at an abstract level in the ISO standard for an Open Archival Information System (OAIS) Reference Model.¹ The OAIS functional model divides the lifecycle in seven main functional entities: Ingest, Archival Storage, Data Management, Preservation Planning, Administration, Access and Common Services. The entities are each comprised of several functions described in detail in the standard. Organisations dedicated to long-term preservation are termed trustworthy digital repositories.²

In practice repositories are implemented in various ways according to on the one hand, the available resources, and on the other, the nature of the organisation and the assets in its trust. The repository may, for example, be private or public, it may have a business need or a legal mandate to preserve, and the duration of preservation may be definite or infinite. The assets, whether individual objects, or groups of objects, may represent documents, books, images, films, web pages, etc., in various qualities. Here quality is used as a collective name for different properties of the assets relating to their content, context, appearance, structure and behaviour³ as well information security,⁴ including authenticity, confidentiality, integrity and availability. On a more detailed level quality also relate to format types, a book may, for example, be saved as a pdf or word document, or as tiff images; preservation strategies, whether based on migration or emulation; and how the repository is organised, e.g., as a distributed or centralised system.

In order to allocate sufficient resources for digital preservation activities, repository managers obviously need to know how much they cost. However, as may be perceived from the above examples of the diversity in the digital preservation scenarios and the comprehensiveness of preservation activities in depth as well as in breadth, it is a complex task to account for the costs and the cost drivers of digital preservation. Furthermore, as opposed to traditional preservation of, for example, paper-based records, digital preservation services have not yet fully matured. This means that costing of digital preservation is not only complicated by constantly evolving technologies, but also because best practice has not been established.

Over the last decade several cost models have emerged to support cost assessment, some aimed at specific sectors or materials, others more generic in nature: The National Archives of the Netherlands' cost model for digital preservation,⁵ the NASA Cost Estimation Tool (CET),⁶ the LIFE Costing Model,⁷

¹ Consultative Committee for Space Data Systems (CCSDS), *Reference Model for an Open Archival Information System (OAIS)*, Magenta Book, Recommended Practice, CCSDS 650.0-M-2, 2012 (ISO14721:2003), accessed August 2012, <http://public.ccsds.org/publications/archive/650x0m2.pdf>.

² Consultative Committee for Space Data Systems (CCSDS), *Audit and Certification of Trustworthy Digital Repositories*. CCSDS 652.0-M-1, Magenta Book, 2011, (ISO 16363:2012), accessed August 2012, <http://public.ccsds.org/publications/archive/652x0m1.pdf>.

³ The InSPECT (Investigating the Significant Properties of Electronic Content Over Time) project, "Significant Properties Report," 8, 2009, accessed August 2012, http://www.significantproperties.org.uk/wp22_significant_properties.pdf.

⁴ ISO/IEC 2700:2009, Information technology — Security techniques — Information security management systems — Overview and vocabulary, 3, 2009.

⁵ Slats, J., and Verdegem, R., "Cost Model for Digital Preservation," in *Proceedings of the IVth triennial conference, DLM Forum, Archive, Records and Information Management in Europe*, 2005, accessed August 2012, http://dmlforum.typepad.com/Paper_RemcoVerdegem_and_JS_CostModelfordigitalpreservation.pdf.

⁶ NASA, "Cost Estimation Toolkit (CET)," accessed August 2012, <http://opensource.gsfc.nasa.gov/projects/CET/index.php>.

the Keeping Research Data Safe (KRDS) model,⁸ the Cost Model for Digital Archiving,⁹ the Cost Model for Digital Preservation (CMDP),¹⁰ the DP4lib cost model,¹¹ the PrestoPRIME cost modelling tools,¹² the California Digital Library cost model,¹³ and the Economic Model of Storage.¹⁴

The costs of the preservation activities cannot be viewed in isolation, but need to be balanced against the benefits that they represent to stakeholders. These aspects of the economies of digital preservation were investigated by the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, and resulted in two comprehensive reports^{15,16} and initiated the development of the Economic Sustainability Reference Model.¹⁷ To this end the KRDS project has also developed tools for analysing benefits¹⁸ and in the Cost Model for Digital Archiving benefits are expressed via the Balanced Scorecard methodology.¹⁹ Strategies for funding digital preservation are also described in a recent report on alignment in digital preservation initiatives.²⁰

Accounting for the costs of preservation activities, their relationships, variables, and underlying pre-conditions in a structured manner remains highly challenging and so far no consensus has been reached within the community on how to model the costs of preserving digital assets.

In this paper we examine which cost data are needed for managing a repository, and how the costs of preserving access to digital assets are influenced by other economic factors. Based on this examination

⁷ Hole, B., Lin, L., McCann, P., and Wheatley, P., "LIFE3: A Predictive Costing Tool for Digital Collections," in *Proceedings of iPRES 2010, 7th International Conference on Preservation of Digital Objects*, Austria, 2010, accessed August 2012, <http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hole-64.pdf>.

⁸ N. Beagrie, B. Lavoie, and M. Woollard, *Keeping Research Data Safe 2, Final Report*, Charles Beagrie Ltd., 2010, accessed August 2012, www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf.

⁹ A. S. Palaialogk, A. A. Economides, H. D. Tjalsma, and L. B. Sesink, "An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS," *Int J Digit Libr*, accessed August 2012, doi:10.1007/s00799-012-0092-1, Springer, 2012, <http://rd.springer.com/article/10.1007/s00799-012-0092-1>.

¹⁰ U. B. Kejser, A. B. Nielsen, and A. Thirifays, "Cost Model for Digital Preservation: Cost of Digital Migration," *The International Journal of Digital Curation* 6, no. 1 (2011): 255-267, accessed August 2012, www.ijdc.net/index.php/ijdc/article/view/177.

¹¹ DP4lib Kostenmodell für Langzeitarchivierung, http://dp4lib.langzeitarchivierung.de/downloads/DP4lib-Kostenmodell_eines_LZA-Dienstes_v1.0.pdf.

¹² M. Addis, and M. Jacyno, "Tools for modelling and simulating migration based preservation," PrestoPRIME, 2010, accessed August 2012, https://prestoprime.ina.fr/public/deliverables/PP_WP2_D2.1.2_-_PreservationModellingTools_R0_v1.00.pdf.

¹³ California Digital Library, CDL, "Cost Modeling," <https://wiki.ucop.edu/display/Curation/Cost+Modeling>.

¹⁴ D. Rosenthal, "Economic model of Storage," November 2011, accessed August 2012, <http://blog.dshr.org/2011/11/progress-on-economic-model-of-storage.html>.

¹⁵ Blue Ribbon Task Force on Sustainable Digital Preservation and Access, "Sustainable Economics for a Digital Plant: Ensuring Long-Term Access to Digital Information," Final report, 2010, accessed August 2012, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

¹⁶ Blue Ribbon Task Force on Sustainable Digital Preservation and Access, "Sustaining the digital investment: Issues and challenges of economically sustainable digital preservation," Interim report, 2008, accessed August 2012, http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf.

¹⁷ C. Rusbridge, "Update on the state of the Economic Sustainability Reference Model," accessed August 2012, <https://unsustainableideas.wordpress.com/2011/10/17/update-state-ref-model/>.

¹⁸ C. Beagrie, "KRDS/I2S2 Digital Preservation Benefit Analysis Tools Project," accessed August 2012, <http://beagrie.com/krds-i2s2.php>.

¹⁹ R. S. Kaplan, and D. P. Norton, *The Balanced Scorecard: Translating Strategy into Action* (Boston: Harvard Business School Press, 1996).

²⁰ M. Lunghi, N. Grindley, B. Stoklasová, A. Trehub, and C. Egger, "Economic Alignment," in *Aligning National Approaches to Digital Preservation*, ed. N. McGovern (Educopia Institute Publication, 2012), 195-268, accessed August 2012, <http://educopia.org/publications/ANADP>.

and with a view to existing cost methodologies we analyse the constituents of a generic cost model, capable of accounting for different types of assets and organisations. This work is based on experience from the Danish CMDP project (2008-2012) in which a cost model and an associated spreadsheet tool for assessment of the cost of preserving cultural heritage materials have been developed.²¹ Wherever relevant the CMDP model and tool have been used to exemplify how different challenges in cost modelling may be addressed.

2. Required Preservation Cost Data

2.1 Budgeting

Cost data for digital preservation is needed for preparing and controlling budgets. Repository managers need to know the total costs of all activities involved in the digital preservation lifecycle and the distribution of costs over individual activities, and groups of activities. They must also be aware of which activities are included in the preservation lifecycle and which are not. Understanding the distribution of costs is also important for identifying when the costs of specific activities fall due, e.g., when re-investments in storage media or systems are required, and for pricing specific preservation services in relation to outsourcing. Likewise, cost data is needed for evaluating the effect of adjusting the cost of one lifecycle activity on the other activities, e.g., if the richness of metadata provided at ingest is decreased to save costs, it may induce increased costs at access. In addition, it is needed for identifying the most important costs, which in particular require careful monitoring.

Managers must also know the accounting principles underlying cost figures. For example they need to know if the cost includes the full economic costs, i.e., the direct investment and operation costs, as well as indirect costs, such as the cost of general administration and facilities (overhead). If indirect costs are included they must in addition know how they have been distributed, e.g., as a percentage over all lifecycle activities, or on individual activities. Managers also need to know how possible financial adjustments, e.g., inflation, are accounted for. Likewise, they must acknowledge the assumptions on which estimates of future costs are based, e.g., are assumptions based on projections of historical data or is status quo assumed in price development.

2.2 Optimisation

Cost data is also required for adjusting expenditures to the projected financing and for potential optimisation of preservation activities. Within an existing budget framework it may be possible to enhance the preservation systems and processes, without compromising the quality, and thereby improving the overall cost efficiency of the preservation activities. This may, for example, be achieved by engaging in partnerships that allow exploiting economies of scale, e.g., in the area of archival storage, or economies of scope, e.g., by providing more versions of the same asset. Going beyond the existing framework, budgets may be balanced by attracting additional funding from external stakeholders or by internal re-allocation of resources within the organisation. It is also possible to adjust budgets by reconsidering the quantity and/or the level of quality of the assets to be preserved.

²¹ “The Cost Model for Digital Preservation,” accessed August 2012, www.costmodelfordigitalpreservation.dk.

2.3 Evaluation of alternatives

Cost data is also needed for comparing the costs of alternatives, e.g., the costs of applying different preservation strategies or the cost of increasing certain quality properties of the assets. In order to execute such comparisons managers need to be able to account for the quality of the preservation activities. If, for example, the cost of migration processes are compared it is necessary to specify, among other things, how well significant properties of the assets are preserved; how many errors in the process are acceptable and how many random samples are needed to detect errors.

2.4 Economic aspects

The nature of the assets and the organisations that are responsible for preservation and access motivate stakeholders', i.e., owners, producers, and consumers of the assets, requirements for the quality of the assets. Based on these requirements the repository managers specify the quality of the assets. This value proposition represents a certain value to stakeholders and determines their willingness to pay for the assets (see Figure 1).

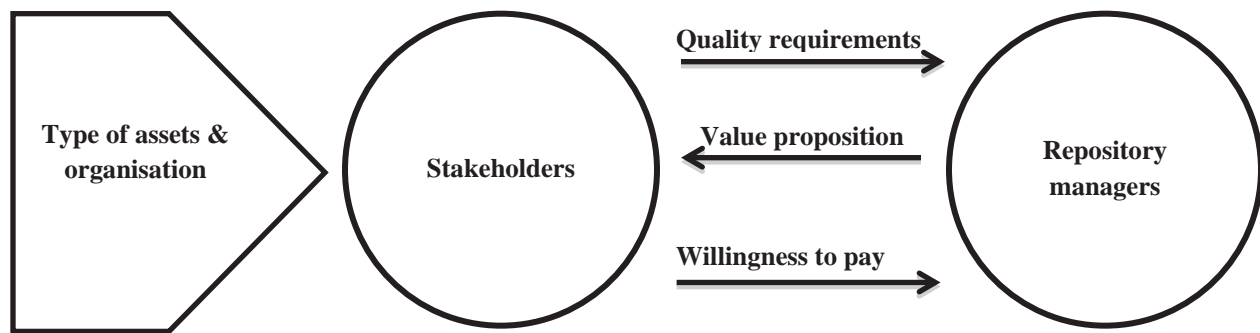


Figure 1. Interactions between stakeholders and repository managers.

Thus, the costs of the preservation activities and the associated quality of the digital assets and services provided by the repository must be seen in relation to the perceived value and stakeholders' willingness to pay for the assets. Whenever changes in the quality of the assets occur they are likely to influence how stakeholders value the preserved assets and thereby their willingness to pay. A digitized book may, for example, gain value if the file is processed by an OCR program to make the text searchable. Likewise, making a repository's degree of trustworthiness explicit by investing in audit and certification could potentially increase revenues for a preservation service provider. Therefore managers must also be aware of the business models underlying the repository, and understand the mechanisms by which the value of assets and preservation services is driven.

2.5 Towards a generic cost model

In order to optimise digital preservation activities and identify best practices it is important to account for the costs and the quality of the activities in a complete and consistent way that further allows for comparing different scenarios.

Ideally, a generic cost model should apply the same methodology for accounting for the costs of all types of assets, preserved in different qualities, and by different preservation strategies, services and organisations; and it should therefore describe the required preservation activities at a conceptual and implementation neutral level. However, costs cannot be assigned on the conceptual level; it is only possible to cost actual implementations. The reason is that activities can be implemented in several ways resulting in different qualities and costs. Thus it is prerequisite to specify the actual implementations in order to assign resources to it, and cost assessments are therefore based on a series of assumptions about the quality of the assets and how they are preserved.

In the CMDP project there is a loose distinction between a cost model and a tool: The CMDP model defines the principles and methods for assessing costs of preserving assets, whereas the CMDP tool makes these principles operational and enables actual cost assessment in a spreadsheet.

A generic cost model should also determine the drivers of the costs in digital preservation. In addition, it should describe the principles for accounting, making financial adjustments, and projecting future costs.

Based on the examination of cost data required for costing digital preservation we propose that a cost model should comprise four building blocks: Profiles, Preservation activities, Cost drivers and Accounting principles (see Figure 2). These model components are described in more detail in the following sections.

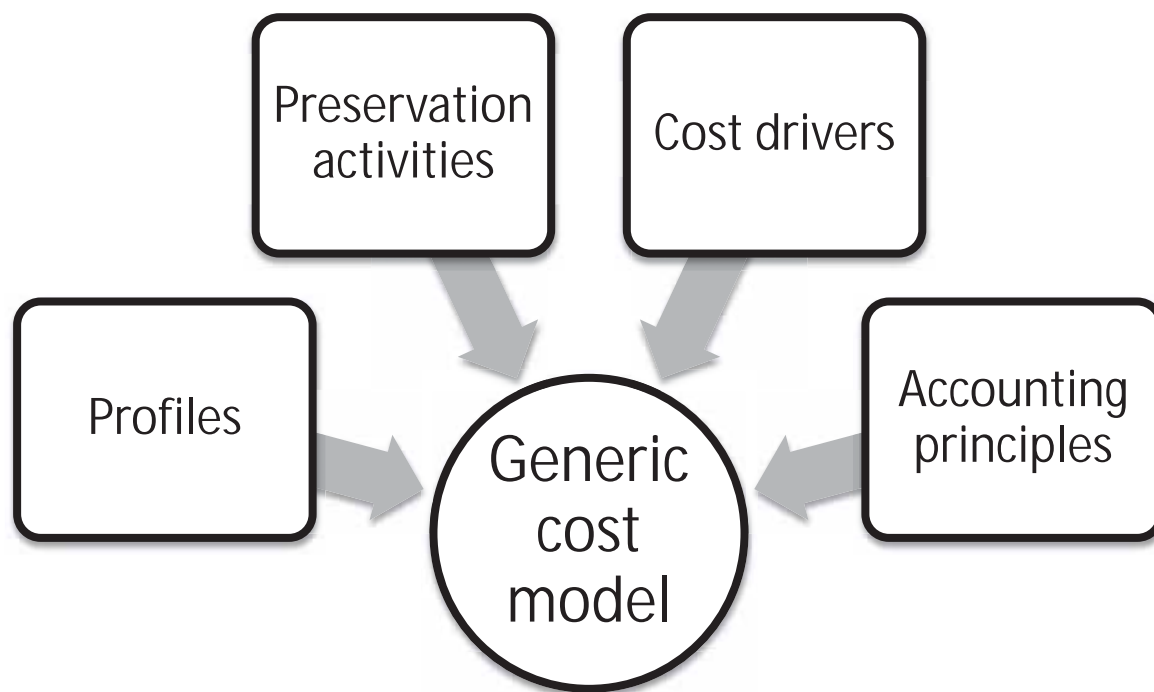


Figure 2. Proposed building blocks of a generic cost model

2.6 Profiles

One way to map between the conceptual cost model level and specific preservation scenarios is through the use of profiles that define the context of the cost assessment. For example a profile could represent an archive with a legal mandate to preserve cultural heritage documents or a private service provider offering

cloud storage. Well-described profiles could help understanding and interpreting cost figures and support cost comparisons.

In the CMDP tool it is possible to add contextual information via predefined dropdown lists where users can select, among other things, the type of organisation, e.g., archive, library; the purpose of preservation, e.g., business requirement, legal mandate; and potentially also quality requirements, e.g., regarding the confidentiality of the assets. The idea is to gradually develop a database with specific cost profiles for different scenarios, which can then be used for comparing sets of cost data.

2.7 Preservation activities

To ensure that the relevant costs are accounted for, the models must identify, describe and delimit all major activities in digital preservation in a complete, precise and systematic way; and describe the relationships and dependencies between the activities.

From its inception the OAIS Reference Model was proposed as the basis for costing the activities in digital preservation,²² and most cost models, including the CMDP, are more or less strictly based on the OAIS functional model. The OAIS functions are described at an abstract level that must be transformed into specific activities, which can then be allocated resources. For example, the OAIS standard states that the function Generate AIP under the functional entity Ingest “may involve file format conversions.” Before you can assess the costs of this function, you need to make a series of assumptions regarding the assets to be preserved, the applied preservation strategy, the technical and organisational design of the repository, and the required quality of the conversion.²³

In the CMDP project the functional descriptions in the OAIS standard and the flow between the functions were analysed to identify “cost-critical” activities, i.e., tasks, which take more than one-person-week to accomplish. Resources were then associated with the activities and cost dependencies identified. The resources and relationships were then expressed in mathematical formulas and made operational in the CMDP tool. So far expressions for the functional entities Ingest, Archival Storage, Preservation Planning, and Administration have been completed and implemented in the tool. The tool works by summing up the cost of the activities, functions and functional entities over the costing period.

2.8 Accounting principles

A generic cost model must define the applied accounting principles, i.e., the standards, rules and procedures used for recording and reporting accounting information to obtain consistency and accuracy in financial statements. The model should also state how possible financial adjustments, e.g., to account for inflation, are made, and how investments are depreciated.

As do most current cost models, the CMDP tool applies activity based costing,²⁴ and it refers to the International Cost Model Standard.²⁵ The tool breaks the cost of individual activities down in the

²² S. Sanett, “Toward developing a framework of cost elements for preserving authentic electronic records into perpetuity,” *College & Research Libraries* 63, no. 5 (2002): 388-404.

²³ U. B. Kejser, A. B. Nielsen, and A. Thirifays, “Cost Aspects of Ingest and Normalization,” in *Proceedings of the iPRES2011 Conference*, November 1-4, 2011, Singapore, pp. 1-10.

²⁴ R. Cooper, R. S. Kaplan, L. S. Maisel, E. Morrissey, and R. M. Oehm, *Implementing Activity-Based Cost Management: Moving from Analysis to Action* (New Jersey: Montvale, Institute of Management Accountants, 1992).

effective time required to complete the activity, measured in person-weeks, multiplied by the salary level, plus purchases, i.e., monetary value. It includes all direct expenses from investment, operation and maintenance, but it does not yet account for indirect costs (overhead). Depending on what type of cost analysis is needed, all investment costs can be depreciated linearly over their useful lifetime or they can be accounted for in the year that they fall due. The CMDP tool applies three salary levels, which are: manager, computer scientist, and technician. As all constants in the CMDP tool the actual salary levels can be individually adjusted to account for specific cases.

Predictions of preservation costs beyond just a few years are highly uncertain increasing exponentially over the years. This uncertainty is especially due to the difficulty of predicting future costs of technology. Future costs estimates can be based on projecting historic cost data and either assuming status quo or an increase or decrease in costs. However, uncontrollable incidents may occur that radically change the former course of the costs. For example, the cost of storage media per capacity has decreased over the last decades, but the flooding catastrophe in 2010 in Thailand, suddenly made prices increase significantly on the short term. Likewise, economic crises, such as in 2008, are likely to influence the rate of increase in costs of, for example, salaries. In addition, conflicting tendencies add to the uncertainty, e.g., on the one hand formats tend to become more complex, which, if all else is equal, will increase the costs of managing them; on the other hand, it has also been argued that formats tend to become more stable, decreasing the needed migration frequency in a migration strategy, and thus the costs.

Currently the CMDP tool does not account for financial adjustments. Ideally these should be included, but with a view to the overall uncertainty of the cost projections, this has not yet been given priority. Due to the lack of more accurate estimates, the CMDP tool assumes that costs remain status quo.

2.9 Cost drivers

The total costs of preserving assets are comprised of fixed costs and variable costs. The latter depend on the quantity and quality of assets and the expected duration of preservation, whereas the fixed costs, e.g., salaries, rents and utilities, are relatively independent of these, until a certain productivity level. Thus there are some minimum requirements for operating a repository that must be fulfilled; it must, for example, have facilities, systems and people in place, and these fixed costs can provide for a certain production of assets. If the repository does not exploit its capacity, the costs per assets are higher than if it operated at full capacity.

The quantity of assets to be preserved and their size (data volume) is obviously an important driver of costs, and especially storage costs.²⁶ In the CMDP tool quantity is accounted for by the annual increase in the amount of data (GB) to be preserved over the costing period. It is not yet possible to also enter the number of assets, which can in some cases counter that large files are disproportionately more expensive than small files.

A generic cost model must enable accounting for the quality of the preserved assets. The quality of the assets, including the type of format and how they are preserved, influences the preservation costs at different levels: From the lowest level representing the quality of the individual preservation activities, to the quality of groups of activities to the quality of the whole preservation lifecycle. The quality of the

²⁵ OECD, "International Standard Cost Model Manual to reduce administrative burdens," 2004, accessed August 2012, www.oecd.org/dataoecd/32/54/34227698.pdf.

²⁶ A. B. Nielsen, A. Thirifays, and U. B. Kejser, "Costs of Archival Storage," in *Proceedings of the Archiving 2012 Conference*, 2012, pp. 205-210.

format, e.g., its complexity and whether it is compressed or not, may influence the data volume, and thus the storage costs, but it may also affect the cost of managing the assets over time; all things being equal, simple formats preserved by the migration strategy tend to be less expensive to manage over time than more complex formats. One of the most important cost drivers here is the migration frequency and thus the expected longevity of formats. In the end the overall quality of the preservation activities defines the value that the assets accrue for stakeholders.

In the CMDP tool quality is only accounted for implicitly by the selections that users of the cost tool make when inputting data. Thus, the quality of the information security is, for example, partly defined by the selected number of instances (copies) of each asset and by the selected types of preservation systems, e.g., based on optical or magnetic media storage media, on-line, near-line or off-line solutions.

Likewise, the quality of the assets is partly defined by the selected types of formats and their ability to preserve significant properties. Currently, the CMDP tool assumes the use of the migration strategy with normalisation at ingest. It is possible to select the type of production format from a list and, if required, the tool also suggests suitable preservation formats. The tool can also account for the cost of ingesting the assets in their original format and keeping them status quo for a future emulation. However, it cannot estimate for the future costs of emulation tools to make the assets accessible if the original format has become obsolete. The tool predicts the longevity of the formats individually, but these pre-defined values can be changed.

The perceived duration of preservation is also a significant factor in cost assessment. The CMDP tool allows users to estimate cost over 20 years. In relation to the before mentioned difficulties in projecting cost, this is of course a very uncertain prediction. However, such estimates may still be of use because they show tendencies in cost projections and indicate when important costs are expected to fall due.

3. Discussion

One may ask if it is realistic to aim for a generic cost model capable of accounting for all possible preservation scenarios and use cases, including budgeting, optimisation, and comparing alternatives. Based on experiences from the CMDP project it seems possible to design a generic cost model for as long as the model is kept at a sufficiently conceptual level. The challenges rise however, on implementing the general principles in a tool that can make the actual cost assessments. Creating cost profiles for specific scenarios may be a way to meet this challenge. The profiles should include contextual descriptions of the nature of the assets and the organisation responsible for preserving the assets, and of the resulting requirements for the quality of the assets.

Archival storage represents the least complex and most mature area within the digital preservation lifecycle. It has received considerable attention with regards to cost modelling and it is the area from which most empirical cost data is available. Therefore, the developers of a generic cost model would likely benefit from using archival storage as the first area to develop, and then use the achieved results to leverage the remaining areas of the lifecycle.

Defining a common terminology related to costs and benefits of preserving digital assets is an important prerequisite for agreeing on requirements for a conceptual cost model. The terminology provided by the OAIS standard and other vocabularies developed within the digital preservation

community constitute a good basis, but they need to be extended with definitions of economic concepts and a business-oriented vocabulary.

The OAIS functional model provides a sound basis for accounting for cost of preserving digital assets, since it describes all major functions in digital preservation. Furthermore, it is already used for benchmarking organisations, e.g., in relation to audit and certification.²⁷ Cost models based on OAIS are structured accordingly, and other cost perspectives may be required, e.g., ones oriented at specific business processes.²⁸ However, if the cost methodology is sufficiently detailed and well-documented, it should be possible to assemble the activities deducted from the OAIS functional descriptions in groups that allow for analysing other specific processes.

The complexity of the digital preservation landscape entails very detailed cost models that are difficult to understand and use by anyone other than their own designers; on the other hand simpler models are often not accurate and precise enough to provide the required cost data. Thus there is a need for addressing the challenge between the complexity of the models and their user-friendliness.

One of the most difficult things in cost modelling is to identify and account for cost drivers. Especially, it is not straightforward to model how the various aspects of quality of the preserved assets influence costs. Initial work has been done to express certain quality aspects of information security mathematically,²⁹ but so far quality can only be accounted for in relatively subjective terms. At a high level, one way forward in benchmarking the quality of the assets could be to associate the costs of preservation with the certification rankings obtained by a repository through audit.

Assessment of quality is further complicated by the fact that the perceived value of a certain quality aspect is not an absolute measure, but dependent on the benefits that it represents to specific stakeholders. Thus, there is a need for combining cost models with benefit and business models to account for other economic factors.

4. Conclusion

Preservation cost data constitutes an important source for managing repositories, and cost data is required for various purposes including budgeting, optimisation, and evaluation of alternative preservation opportunities. They are also essential for the assessment of the value of the preserved assets and thereby stakeholders' willingness to pay for this value. The increasing amount of digital information to be preserved underlines the need for optimising the cost-efficiency of preservation activities and reaching best practices within the community. This requires complete and consistently collected cost data, which can be achieved through the use of a generic cost model.

Developers of cost models should therefore cooperate to develop a generic cost model to which profiles for specific preservation scenarios could be applied. Having created such a common framework it will be possible and necessary to have well-documented cost data to test and improve the coverage and

²⁷ ISO 16363:2012, "Space data and information transfer systems -- Audit and certification of trustworthy digital repositories," 2012.

²⁸ "COBIT 5: A Business Framework for the Governance and Management of Enterprise IT," accessed August 2012, <http://www.isaca.org/cobit/pages/default.aspx>.

²⁹ E. Zierau, U. B. Kejser, and H. Kulovits, "Evaluation of Bit Preservation Strategies," in *Proceedings of the 7th International Conference on Preservation of Digital Objects (iPRES), Vienna, Austria, 2010*, pp. 161-169, accessed August 2012, <http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-31.pdf>.

precision of existing models and tools. However, this is hard task since digital preservation is not yet a mature business with a best practice.

In order to allow for comparing the costs of preserving assets for different organisations it is also important to account for the quality related to the preserved assets and the value that it represent to stakeholders. Therefore, it is important to link cost models with economic models to represent both cost and benefits of preserving access to digital assets.

Acknowledgements

We wish to thank the Danish Ministry of Culture, The Danish National Archives and Royal Library for their support of the CMDP project.

Economically Easy Method to Digitize Oversized Documents with Special Reference to Ola Leaf Manuscripts in Sri Lanka

L.M. Udaya Prasad Cabral

Head, Conservation & Preservation Division, National Library and Documentation Services Board, Sri Lanka, ubk@kb.dk

Abstract

Ola leaf manuscripts are the native writing medium of Sri Lanka. It had been prepared from the leaves of a tree commonly found in the country. Due to its location, Sri Lanka has a tropical climate. High humidity and temperature have badly damaged many Ola leaf manuscripts still deposited at ancient temples in the country. Preserving, obtaining information of and disseminating the knowledge contained in Ola leaf manuscripts are some key duties of the National Library. The National Library has adopted a procedure to digitize these oversized, uneven and brittle manuscripts for studying, as well as other digitization methods. Digitizing using a 50mm camera with copy table is very economical when compared with other methods. Sharpness of the pictures and three dimensional views provide natural attraction of digital files which take readers as close as possible to the originals. A specially prepared database offers a reader-friendly environment for those interested to gather knowledge from these native writings.

Author

Mr. L.M. Udaya Prasad Cabral is a Postgraduate Scholar. For his higher studies, he is carrying out research on protecting the archival materials from harmful environmental parameters of the tropical region. Study on preservation aspects of ancient library materials of Sri Lanka, Ola leaf manuscripts, is one of his fields of keen interest. Last year he was able to test a herbal extraction that can be used to preserve wooden objects as well as Ola leaf manuscripts. The National Research Institute of Cultural Heritage of Korea assisted him last year in his research through the Asian Cooperation Programme on Conservation Science. Presently he is working as Document Conservator and is the Head of the Conservation and Preservation Division of the National Library of Sri Lanka. The National Library of Sri Lanka acts as a convenor of the MOW programme in Sri Lanka. He provides his full cooperation to promote the MOW projects in the country, as the MOW programme is handled by his division.

1. Introduction

Sri Lanka is a country with glorious history and enriched with precious documentary heritage. Because the National Library is a guardian of the rich documentary heritage of the country, it has a keen responsibility to protect such heritage in its original format as well as disseminate knowledge inscribed in the documentary heritage among the country's younger generations. Ola leaf manuscripts are one of the native writing mediums that our ancestors used to communicate their thoughts and keep records in Sri Lanka's ancient society. These are basically prepared by pre-treating leaves of the *Corypha umbraculifera* tree, which was commonly found in Sri Lanka at that time. Ola leaf manuscripts have been a very popular writing medium since the 12th century A.D. It continued to be the main medium until paper was introduced by the Dutch Colonialists during the 17th Century.

According to historians, the 3rd century B.C. to 12 century A.D. was the golden period of Sri Lankan history. Our ancestors constructed huge pagodas, irrigation systems, storied buildings, reservoirs as big as the seas, etc., during that period. Modern technology couldn't complete such constructions even

today. Various other fields such as medicine, religious philosophy, etc., were also in a highly developed stage at that time.

Ola leaf manuscripts were the only writing medium at that time. Hence, knowledge on science, technology, religion, philosophy, astrology, medicine and other areas of knowledge were inscribed on Ola leaf manuscripts by our ancestors. Ancient temples became the guardians of these documents when foreign invaders attacked our country on several occasions. Our ancient society had a close connection with temples. They were also education centres. These temples generated, disseminated and deposited the knowledge. These were one of key duties of the temples, in addition to religious activities. Scholars gathered around the temples and these temples automatically became knowledge depository centres. Ola leaves had been the only writing medium, which had been passing indigenous knowledge from generation to generation. A considerable number of Ola leaf manuscripts have been protected in ancient temples until now. Ola leaf manuscripts contain invaluable information. Some Ola leaf manuscripts, especially on Buddhist religious teachings and practices, have been printed as books, but there are many manuscripts locked in almirahs at ancient temples that have not been read by anybody. Since monks believe these manuscripts reveal the route to treasures hidden by past kings, nobody gets permission to unlock the almirahs to read manuscripts for any purpose.

Sri Lanka is a tropical island located in the Indian Ocean just below the Indian peninsula. Its location is responsible for the high humidity and high temperature climatic conditions experienced in the country throughout the year. Ola leaf manuscripts deposited in the temples and libraries have been affected by these climatic conditions. Environmental parameters and unsuitable storage conditions have created conservation problems and accelerated the deterioration process of the Ola leaf manuscripts. It is obvious that almost all the Ola leaf manuscripts, which are considered as treasures of the documentary heritage of this country, are in danger, unless the National Library takes comprehensive steps, immediately, to rectify this problem.

The following conservation problems have been identified in the Ola leaf Manuscripts Collection Program, in substance:

- *Leaves sticking together* - Some leaves of the Ola books tend to stick together. Excess oil on the leaves is one reason for this, as the oil, in high temperature conditions, results in this sticky condition.
- *Fading letters* - The writings inscribed on Ola leaf manuscripts have become faded due to environmental conditions
- *Loss of flexibility* - This is due to loss of oil on the Ola leaves. At very low RH levels and in high humidity conditions, the leaves dry out and become hard. Such dry leaves can easily be broken into pieces.
- *Pest attacks* - Ola leaves are easily damaged by insects. The insects make holes through the leaves, rendering them unreadable as well as accelerating their deterioration.
- *Microbial attacks* - Fungi and bacteria have been found and isolated on the manuscripts. Favorable humidity and temperature for the microbes and nutrients substrate provided by the Ola leaf itself create this condition.
- *Discoloration* - Ola leaves have a tendency to become stained because of bio-deterioration.

Several collections of Ola leaf manuscripts are deposited at Depository Libraries in the country. The National Library, as the major depository library in Sri Lanka, has a huge collection of ancient Ola leaves. The Conservation and Preservation Division of NLDSB focuses its attention on preserving them as it is a valuable collection of the library.

2. Preserving Ola leaf manuscripts in two ways

The inimitable knowledge on medicine, science, technology and etc., contained in the ancient Ola leaf manuscripts of Sri Lanka might be lost to the world forever if a comprehensive program for preserving Ola leaf manuscripts is not commenced soon. The National Library is planning the preservation process to achieve two goals: to understand the value of the Ola leaf manuscripts preserved in their original form, and to make arrangements for preserving them using chemical and traditional treatments. On the other hand, they understand the value of the knowledge itself contained in the manuscripts, which is useful for the development of society and to satisfy the thirst for knowledge in our society. To achieve these goals, a newly defined digitization process has been introduced. It is expected that this method will fulfill readers' and researchers' eagerness for easy access to the manuscripts and save the readers valuable time, while helping to popularize these native manuscripts in the society much more than at present. Because physically handling and reading is harmful to the Ola leaves, as they are brittle, the digitization process was proposed as a suitable preservation method. One of the long-term expectations is to extract the scientific, medical, technological, etc., knowledge contained in these manuscripts using the facts on the manuscripts through comprehensive research.

3. Preserving the manuscripts in original forms

Two traditional herbal extractions were tested to preserve the Ola leaf manuscripts in their original form.

3.1 Experimental Method

This study investigated microbial and insecticidal activation against the two herbal extractions named as NL and DNA believed to have been used by our ancestors.

3.2 Extracted samples

NL and DNA herbal emulsions were extracted.

3.3 Strains and insects

The fungi strains¹ (mould) *Cladosporium cladosporioides* (H₁), *Aspergillus sydowii* (H₂), *Penicillium citreonigrum* (H₃), *Penicillium toxicarium* (H₄), *Penicillium corylophilu* (H₅) and *Alternaria spp.* (H₇) commonly found in paper materials were obtained from the micro lab of the National Research Institute of Cultural Heritage (NIRCH) in South Korea. Bacteria strains that were isolated from the ancient Ola

¹ A group of organisms of the same species, having distinctive characteristics but not usually considered a separate breed or variety.

leaves and fresh, untreated Ola leaves used in this experiment. For the analysis of insecticidal activities, *Lasioderma serricorne* (Cigarette beetles), which were bred in NIRCH bio-lab, were used. *Lasioderma serricorne* is the most dangerous library pest found in Sri Lanka.

3.4 Procedure 1

3.4.1 Assessment of Antifungal Activation

Pasteurized paper discs were placed on the cultured plates. The two types of herbal extracts were made to be absorbed in Pasteurized paper disc by 50 μ L by using a paper disc susceptibility measuring method. PDA and cultured mould species (H₁, H₂, H₃, H₄, H₅, H₇) were prepared in the same concentration (3 \times 10⁶ CFU) using spread plate methods. Petri dishes were sealed with sealing tape. Samples were incubated at 28°C for 4 days. Control samples were established adding the same amount of mould species without herbal extracts. Two samples of each batch were prepared. Antifungal activities were observed, resulting in the inhibition zone diameter (GZD).

3.5 Procedure 2

3.5.1 Assessment of Antibacterial Activation

Six bacteria species were isolated from ancient Ola leaves and fresh untreated Ola leaves. The top of each colony was touched with a loop, and the growth was transferred and spread in medium of *Luria bertani* in aseptic condition. The two types of herbal extracts were impregnated in Pasteurized paper disc by 50 μ L by using the paper disc susceptibility measuring method. Petri dishes were sealed with sealing tape. Control samples were established adding the same amount of bacteria species without herbal extracts. Samples were then incubated at 28° C for 2 days. Two samples of each batch were prepared. Antibacterial activities were observed, resulting in the inhibition zone diameter (GZD).

3.6 Procedure 3

3.6.1 Assessment of Ant insecticidal Activation

Pasteurized filter papers were fed with 200 μ L of herbal extracts in three concentrations: 0.3gml⁻¹, 0.1gml⁻¹, 0.05gml⁻¹. Twenty species of *Lasioderma serricorne* were positioned in each Petri dish. The filter papers were installed indirectly with test insects in the Petri dishes. Control samples were established feeding 70% Ethanol to the filter paper. These were bred in the incubator at 28° C and 60% RH for three days. The number of dead insects was examined every 24 hour, for three days.

3.7 Results

3.7.1 Anti-Fungal Effect of Herbal Extracts

Herbal extractions of DNA and NL controlled growth of three species of fungi in PDA media, which were incubated strains of (H₁), *Aspergillus sydowii* (H₂), (H₅) and (H₇), and DNA formed growth inhibition zones of 14mm, 14mm, 13mm, 13mm, and NL formed growth inhibition zones of 13mm, 12mm, 13mm, 12mm.

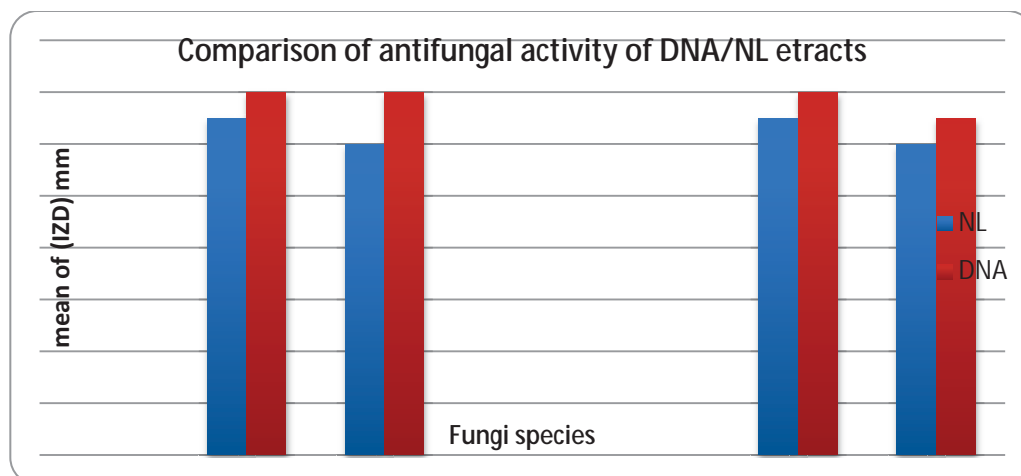


Figure 1. The extraction didn't show antifungal activity against species, (H₃) (H₄).

3.7.2 Anti-Bacterial Effects of Herbal Extracts

Herbal extractions of DNA and NL controlled the growth of species of bacteria that were obtained from the surface of new, untreated as well as ancient Ola leaves. Growth inhibition zones for three bacteria species--(E₆), (E₃) and (E₄)--were obtained from new untreated Ola leaves, while two bacteria species isolated from ancient Ola leaves--(E₂) and (E₅)-- were measured to confirm the antibacterial activity of DNA and NL herbal extracts.

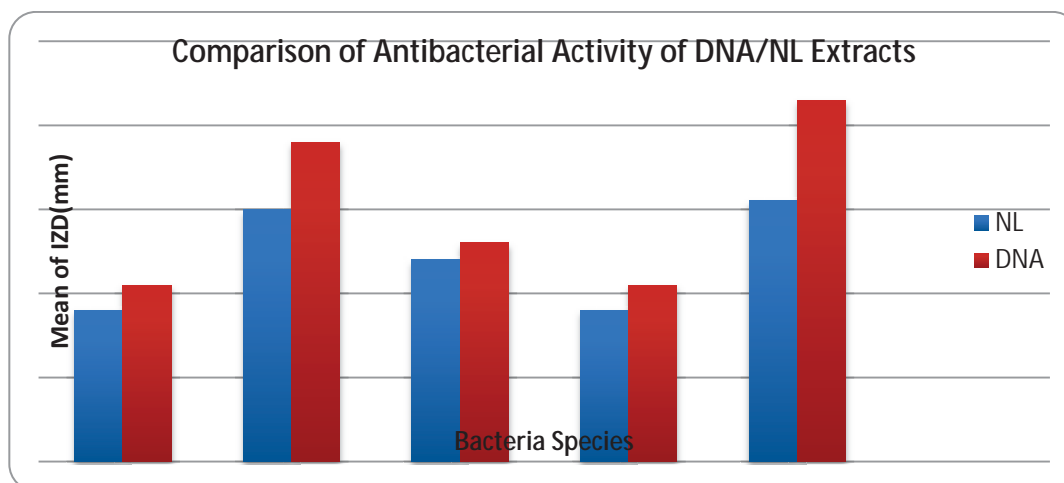


Figure 2. Bacterial strain M₁ didn't show reaction against the herbal extraction.

3.7.3 Insecticidal Effect of Herbal Extracts

Neither DNA nor NL showed any insecticidal effects against *Lasioderma serricorne*. On the third day of the experiment several dead insects were found. These could have been natural deaths. However, insects gathered at one area of the Petri dish and tried to avoid areas of the filter paper, which contained the herbal extract.

Both herbal extractions of DNA and NL showed mild antifungal and antibacterial activities. Results confirmed DNA is more active than NL. NL contains several different ingredients than DNA. Results confirmed that neither herbal extract is highly effective against the insect *Lasioderma serricorne*, but it was observed that most of the insects had gathered in one area of the Petri dishes, which were the extraction-free areas. They may have tried to avoid contacting the herbal extractions. These extractions have some effects against insects. These are not insecticide effects but rather insect repellent effects. The herbal extraction of DNA was selected to use in traditional conservation before digitization.

4. Preserving Manuscripts by Digitization

Digitization is a practical solution that is able to satisfy readers who are interesting in Ola leaf manuscripts. The concept of digitization of library material as a preservation method is a new idea for our library. Actually it is still a developing idea in the library sector in our country. Lack of knowledge about this modern trend of digitization, and budgetary restraints for buying instruments are the main problems confronting us.

Ola leaf manuscripts come in different sizes. Normally, they are 4 to 5 inches wide, but their lengths vary from ½ foot to 3 feet. A manuscript normally consists of 10 – 250 leaves. Even large size scanners available in the country cannot accommodate some of these manuscripts. On the other, hand scanners are very expensive. The A₀ size scanner, which is named the ‘face-up’ scanner in the market, is common in libraries for digitizing books. These are most suitable for the digitization process.

Scanning is an easy and convenient technology that is used for the digitization of books. In this particular situation, however, such scanning technology is not suitable due to several reasons. First, as mentioned, the manuscripts (leaves) vary from ½ foot to more than 3 feet in length. Second, scanning the surfaces alone is not always enough. The nature of the Ola leaf manuscripts is completely different from that of books. For instance, in books printed letters are found on the surface of the paper, but in the case of Ola leaf manuscripts, the letters are engraved on the surface of the leaf; each letter has a depth. The Ola leaf surface is not smooth like paper, and not as clear and even as paper. Third, some Ola leaf manuscripts are very fragile and difficult to handle as they are large sized documents. The A₀ type face-up scanner, attached to two digital cameras is suitable for this purpose. However, several disadvantages are encountered with this instrument. As described earlier, most of the Ola leaf manuscripts are still found deposited in ancient temples in various part of the country. Digitization should be carried out at the respective temple premises. Handling a face-up scanner in a temple, which has limited facilities, is a very tiresome job. Several workers are required to perform this task. Transportation is also a difficult problem. Thus, scanning Ola leaf manuscripts using face-up scanners is a very costly job. We cannot be sure, in advance, that the final picture received after scanning will trace all the characteristics of the manuscripts. It might not provide/indicate depth of the letters as a three dimensional picture of the Ola leaves. Readers prefer to actually see the real Ola leaf manuscripts. So, three dimensional pictures are very important to fulfil the readers’ desires.

Digitization by camera is the best and most efficient method. The cameras are not heavy; one worker can arrange the instruments. In addition, a vehicle for transportation is not necessary. The working area is illuminated by two lights. Normal daylight can also be used.

The National Library of Sri Lanka was able to develop this economically easy, user-friendly method for digitizing oversized Ola leaf manuscripts, after investigating issues regarding digitization of Ola leaf manuscripts and available technologies in the country. This method is the result of a combination

of three fields: traditional conservation treatment methods, photography and computer literacy. This economically easy, user-friendly conservation and preservation method can be described in three steps:

- Step 1 - Brittle Ola leaf manuscripts are treated according to traditional treatment methods.
- Step 2 - Treated Ola leaf manuscripts are digitized using a digital camera.
- Step 3 - Digital data are edited and imported into a database.

4.1 Step 1

Brittle and deteriorated Ola leaf manuscripts are selected from the library for digitization. Letters on the Ola leaves fade when they age. It is very difficult to recognize the letters with normal sight or through a digital camera, which is unable to capture a clear picture without blackening of the letters. A Conservator has to blacken the letters using finely powdered charcoal derived from *Trema orientales*. The National Library of Sri Lanka was able to invent a mixture of herbal extractions that is similar to the extracts used by our ancestors from earliest time for protecting these manuscripts (DNA extraction). This solution reacts on the Ola leaf and protects it in various ways. Herbal extractions give high flexibility to brittle leaves. In addition, applying charcoal powder mixed with herbal extractions gives bright colour to faded letters. Some ingredients of the extract react as fungicides and bactericides. This prevents the growth of microorganisms on the Ola leaves due to high humidity conditions in tropical climates. The smell of the extract works as an insect repellent and protects the leaves from library pests. The adherence property of this extraction is put to good use and preserves Ola leaf manuscripts for a long time.

4.2 Step 2

At the second stage, Ola leaves are photographed using a digital camera. Suitable technical devices and necessary technical support are essential at this stage. Over-size images taken by the camera are not evenly clear and sharp at the edges, especially when the leaf is very long. Sometimes, the image is not bright enough to be easily read. Sometimes the image is not like in the original leaf, most probably because its colour differs from the original leaf.

Selecting a suitable camera is a very important factor in this process. Several technical aspects of cameras were tested. For example, different lens', apertures, shutters speeds and film speeds were investigated. It was determined that a full frame sensor 36X24 camera, which can accommodate a 50mm lens and capture images using more than 20 megapixels, produces a clear, sharp image for the digitization process. Even long leaves (3 feet) can be captured as a clear picture by this camera. The copy stand is an important device that is essential for this process. It provides support to the Ola leaf, which is placed on it parallel to the camera lens situated above it. The camera should always level to the supports where the Ola leaf is accommodated and care should be taken to ensure that the entire document (leaf) flush against copy stand during the entire process. For the best results, the working area should be lighted with two flash bulbs of at least 300W each. For the best results, lights can be softened by umbrellas for each light or by using soft boxes. Diffused light in this direction can be controlled remotely and more effectively by using a radio slave flash trigger. Verifying the colour profile of the computer ensures that the colour displayed on the screen are as close as possible to the colour recorded by the camera. The colour should be adjusted for the largest resolution made on the computer. This is very important.

A particular book of Ola leaf manuscripts consists of a bundle of leaves. Each bundle is photographed and the record is digitally stored as < name of manuscripts cropped > JPEG files.

4.3 Step 3

The computer operator edits the files using Adobe Photoshop and converts the files to pdf format. At the final stage, the digital data are imported into a computer database created by the Information Technology division of the library (Figure 3). Digitization of Ola leaf manuscripts by camera is a very cheap and practical method. This method might be very important for libraries in countries that have traditional library materials similar to Ola leaf manuscripts. It is an obvious solution for institutions that have limited annual budgets and are only able to start a digitization project at the minimal level. Inexpensive manual cameras can even be used instead of digital cameras. This digitization process has a lower budgetary requirement and requires fewer human resources to carry out. Hence, it is a practical and realistic method for developing countries. The high humidity and high temperatures of tropical climates badly affect the manuscripts. Therefore, all the original manuscripts should be stored in a room equipped with a climate control system. This would help to preserve the original manuscripts from harmful environmental parameters as well as the damages caused by physical handling by the readers.

Ola Leaf Manuscripts Collection

Horoscope

Subject Match all words

Medicine

Religious

Literature

Astrology

Folklore

Other

The National Library a guardian of the rich documentary heritage of the country, it has a keen responsibility to protect them as in original format as well as disseminate knowledge inscribed in the documentary heritage among the younger generations in the country. This data base consists of full digital copy of Ola leaf manuscripts deposited in various locations in the country.

Figure 3.

Readers can access the database (Figure 4) and obtain the desired information in a very short time. It saves readers and researchers valuable. Eventually, researchers might even be able to retrieve manuscripts on their desktop at home via the Internet instead of having to visit the library. However, at this stage, the National Library has not provided online access to the database due to laws and ordinances related to intellectual property issues. In the near future, it is hoped that the National Library will be able to provide online searching and retrieval facilities, with some restrictions.

Ola Leaf Manuscripts Collection

Search bar: **GO**

Titile Match all words ☐

Categories: Medicine, Religious, Literature, Astrology, Folklore, Other

<input type="checkbox"/>	Title	අනුරාධපුරය	View
	Location	Gangarama Temple, No.25, Temple Road, Kandy	
	Date of Compilation	1902	
	Original /Copy	Copy	
	Author	Unknown	
	Brief content	Two methods on preparation of horoscope and forecasting	
	Conservation condition	Brittle	

Print

Figure 4.

References

- Krishan, Gopal, *Digital libraries in electronic information era* (New Delhi: Authors Press, 2000).
- Weber, Hartmut, and Marianne Dörr. *Digitization as a method of preservation?* Final report of a working group of the Deutsche Forschungsgemeinschaft (German Research Association). Translated by Andrew Medlicott (Washington, D.C.: Commission on Preservation and Access, 1997).
- Milburn, Ken. *Digital photography bible* (New York: Hungry Minds, 2000).
- Ostrow, Stephen, *Digitizing historical pictorial collections for the Internet* (Washington: Council on Library and Information Resources, 1998). <http://www.clir.org/pubs/reports/ostrow/pub71.html>.
- “Palm leaf manuscripts preservation by digital camera photography demonstration of method” last modified 2012, <http://64.151.72.194/resources/books/agamas/palm-leaf-demo/>.
- Patel, Santosh, *Library nformation preservation and access* (New Delhi: Authors press, 2003).

Preserving Our Heritage

An Independent Advantage

Patricia Liebetrau

Abstract

Significant national funding is not made available for cultural and heritage digitization and preservation purposes in South Africa. Furthermore a lack of Government support for digitization initiatives has resulted in stifled development and a dearth of skills in this area. The situation has worsened in the current economic climate. There is a dire lack of IT skills for libraries and archives with few training courses and other opportunities for continuing professional development. Minimal digital project management expertise currently exists in the country. These and other challenges have contributed to stagnation and project failures. This paper looks at the valuable role played by independent professionals in assisting organisations to move ahead with projects and to overcome or manage these challenges. Very little has been written on this topic, especially in Southern Africa, so this paper draws largely on personal experiences in collaboration with other independents, from a practical perspective.

Author

Patricia Liebetrau is an independent information professional providing consulting services and training for media development in South Africa and beyond into Africa. Prior to this she worked with Digital Innovation South Africa (DISA) (<http://www.disa.ukzn.ac.za>) over a period of 10 years. This innovative project developed an extensive online digital repository of open access resources around South African heritage that assisted new curriculum development and contributed to e-learning and e-research initiatives. Her skills and interests lie in research and implementation of digital technologies in creating information and knowledge resources for libraries, archives and memory organisations. Her area of specialisation is metadata and she was the first metadata librarian in South Africa. Her current focus is on change management and leadership development for the library and archive profession to support academic scholarly endeavors.

1. Introduction

Cultural heritage organisations and archives, including community archives, are increasingly utilising and exploring ways of using evolving digital technologies to enable their collections to be discovered, accessed and utilised in a Web environment. Many of these organisations have extensive, important and interesting physical collections and resources. Digital technologies are enabling these organisations to reach an online audience.

These same technologies provide countries in continents such as Africa with an overdue opportunity to claim their presence in a global environment. Rich but endangered national cultural heritage resources from national archives, Universities, community archives and other organisations can be digitized and made available for research purposes and preserved for long-term access. The process of digitization itself is a preservation step by minimising handling of precious original resources and artefacts. However once digitized the challenge becomes one of managing preserving data for the long-term.

However several challenges are slowing the rate of digitization efforts in South Africa and in Africa as a whole. Additionally, the fragility of digitized resources can pose a real and underestimated danger.

Digital chaos, resulting from a lack of skills, knowledge and understanding, can only too easily undermine a digitization project. Unsuitable hardware and software purchases, inadequate infrastructure, use of proprietary software with expensive licensing requirements, inconsistent or inappropriate use of international industry standards and other factors impede successful large scale digitisation

Many libraries and archives in Africa have been fortunate in the past in securing funding from philanthropic organisations, private funders and other sources of funding for specific projects. However with the current economic climate, many sources of funding are simply no longer available to assist organisations undertake digitization projects on a large scale. Even more concerning is that sufficient funding is often not available after the end of the project to provide for preservation and ongoing Web access costs. Without strong national support, funding and infrastructure, responsibility falls to individual organisations to provide digitization budgets and cover the associated infrastructure and data preservation costs. In the absence of committed funding or exit strategies, projects may ultimately have to be parked or abandoned. Ultimately the vision is to move from externally funded projects to institutionally committed programmes.

2. South Africa

South Africa, situated at the southern tip of Africa, currently has a population of just less than 60 million people.¹ The country itself is geographically divided into 9 provinces. It is a diverse and multicultural country with 11 official languages although English is widely spoken and considered to be the *de facto* language.

The Internet World Stats website² reports that as of 31 December 2012, South Africa had 6.8 million internet users, just less than 5% of Africa's total internet users. At 140 million estimated internet users in Africa, South Africa is the 5th largest user on the African continent. Low internet usage across the Continent impacts on the ability of Africans to leverage this media to enhance the visibility of African (digital) heritage and academic scholarly output.

3. South African Initiatives

South Africa does not currently enjoy widespread national funding and extensive governmental support for digitization initiatives such as that provided by JISC³ in the United Kingdom. Universities are largely funding their own individual digitization projects, often supplemented by foreign funding, to support e-research and provide global access to their own scholarly resources by building Institutional Repositories (IRs). The National Research Foundation (NRF) has created a National Electronic Theses and Dissertations (NETD) portal⁴ for searching, browsing and accessing South African theses and dissertations from South African Universities that have their repositories open for harvesting. The long-term responsibility of preservation of the data resides within the Universities. A grant from the Carnegie Corporation of New York has enabled the NRF to assist previously disadvantaged Universities with training and hosting of their ETDs on a server housed at the NRF.

¹ <http://www.statssa.gov.za/publications/P0302/P03022011.pdf> (Accessed 26 August 2012).

² <http://www.internetworldstats.com/stats1.htm#africa> (Accessed 26 August 2012).

³ <http://www.jisc.ac.uk/aboutus.aspx> (Accessed 27 August 2012).

⁴ <http://www.netd.ac.za/> (Accessed 27 August 2012).

Under the auspices of the same project, a collaborative publication “Managing Digital Collections: a collaborative initiative on the South African Framework”⁵ was published by the NRF in 2010 as an introductory guide intended to supplement a series of regional training workshops aimed at assisting Universities and heritage organisations gain valuable skills. The regional workshops have, sadly, never taken place.

Several Universities are considering, or are already digitising their own unique resources such as those housed in University Archives and Special Collections. Many of these collections are unique and provide a rich insight into our heritage. The University of Cape Town, for example, has digitized several interesting collections housed in their Manuscripts and Archives Department.⁶ One such project is the digitization of photographs of the San (Bushmen) peoples between 1910 and the late 1920’s. The photographs were taken on numerous expeditions made by Dr Wilhelm Bleek, his sister-in-law Lucy Lloyd and his daughter, Dorothea, to the northern Cape Region of South Africa where the San people lived. This collection, possibly the most unique of UCT Special collections, is listed on UNESCO’s Memory of the World register as being heritage of international importance. The full collection of notebooks, oral histories, drawings and photographs are held across three institutions—South African National Library, the National Gallery and University of Cape Town.

The Campbell Collections of the University of KwaZulu-Natal⁷ is a centre of research excellence with an archive, a museum and a library of rich holdings reflecting the social and cultural heritage of KwaZulu-Natal, South Africa. Several thousand early 20th Century historic photographs from this collection have been digitized and made available online to researchers around the world. The fragile originals are now less frequently handled adding to their longevity. This project was made possible with funding from the Andrew Mellon Foundation.

These and similar projects, although often limited in scope, provide online access to important South African heritage, which would otherwise be little known outside the country. Many of these projects have relied on foreign funding to cover costs of digitization but long-term preservation costs need to be covered by the University itself.

Collaborative digitisation efforts between Universities in South Africa are generally not the norm but two important collaborative initiatives do need mentioning. These are the Digital Innovation in South Africa (DISA)⁸ digital archive of South African socio-political resources and the University of Witwatersrand Rock Art Digital Archive.⁹

3.1 DISA

DISA started as a project, funded by the Andrew Mellon Foundation in 1999. The vision was to digitize and create a freely accessible online scholarly resource focusing on the socio-political history of South Africa, particularly the struggle for freedom during the period from 1950 to 1994. It was a national

⁵ Pat Liebetrau, ed., *Managing Digital Collections: A Collaborative Initiative on the South African Framework* (Pretoria, South Africa: National Research Foundation, 2010).

⁶ Janine Dunlop and Lesley Hart, “Digitisation projects at the University of Cape Town Libraries,” *Innovation* 30 (June 2005): 32-42, accessed 27 August 2012, <http://www.innovation.ukzn.ac.za/InnovationPdfs/No30pp32-42Dunlop&Hart.pdf>.

⁷ <http://campbell.ukzn.ac.za/> (Accessed 27 August 2012).

⁸ <http://www.disa.ukzn.ac.za/> (Accessed 27 August 2012).

⁹ <http://www.sarada.co.za/> (Accessed 27 August 2012).

collaborative project, the first of its kind in the country, partnered by the several South African Universities, The National Archives and The National Library. Selected periodicals were sourced from around the country and missing issues sourced from the USA and the UK. Complete runs of “hard-to-find” anti-apartheid journals were for the first time made available “virtually”, in full text, for a global audience.

The second phase was more ambitious. Archival resources identified by scholars as being important documents in the struggle for freedom were identified from a number of institution’s archival records. These primary source documents came with copyright and privacy issues which were navigated with much paperwork and many hours of correspondence and evoked criticism from some quarters that the resources were “cherry-picked” and were therefore taken out of context. Being funded by an American funder also evoked cries of imperialism.

The funding came to an end in 2009 and without a financial commitment from the partners, a clear exit strategy and on-going funding for preservation of the content, the website and resources have been parked pending negotiations with the National Library of South Africa to maintain the national resource. They will, however, require training and funding in order to undertake the task. The host University declined to maintain the resource for the long-term. The website and content is maintained by volunteers.

Despite criticisms, the project was successful in making available a large number of multimedia resources that have been and are still being extensively used around the world by researchers. What is important in the context of this paper is the significant technological skills development and transfer that took place as a result of the project. This included digital conversion skills, metadata, IT for digital resources and IP for digital libraries and could realistically have created a national platform to provide momentum for digitization efforts. It didn’t. Organisations were “waiting” for “national” policies and strategies and support. The training in partner institutions did however provide the impetus for development of several important heritage projects such as the San photograph digitization project at UCT, mentioned earlier.

3.2 The African Rock Art Digital Archive

The African Rock Art Digital Archive (SARADA) project, the largest of its kind in the world and based at Wits University in Johannesburg, is a “milestone in the digital preservation of Africa and the world’s cultural heritage”¹⁰ and a good example of a national collaborative project. It is funded by The Ringing Rocks Foundation and the Andrew Mellon Foundation and brings together scattered collections into one “virtual” space. Images have been scanned from Museums, Universities and private collections around the country. A significant contribution to the sum of African heritage is being made available online through this project.

It is precisely these kinds of projects that are able to drive cutting edge technological developments and define digital information management boundaries but widespread skills deployment is urgently required to push digitization initiatives forward in South Africa. Valuable skills remain within the project, often invested in short term contract staff and student assistants until project insecurity drives them further afield, taking their skills with them, often outside the country!

¹⁰ <http://www.sarada.co.za> (Accessed 27 August 2012).

4. National Register of Digital Initiatives

Browsing through the NRF register of digitization initiatives in South Africa¹¹ (another project funded by the Carnegie Corporation of New York) indicates that the majority of the initiatives listed on the database are still in the planning stage with just a very few projects in progress or completed. What is holding back the realisation of these initiatives?

5. Public Sector

Examining some of the reasons for the slow uptake of digitization in the Public Sector it is clear that the lack of national leadership, a national policy and a national strategy providing support for a standardised approach to developing digital resources has resulted in organisations “waiting” for national guidance before proceeding. A draft of the long awaited National Policy on the Digitisation of Heritage Resources¹² was made public in 2010 for comment but remains to be published.

Added to this is the lack of national funding to undertake digitization projects and limited training options available for continuing skills development. This has resulted in a dearth of suitable skills and a paralysis in the industry.

6. Private Sector

The situation in the Private Sector is different. The positive effect of a Web presence on scholarly communications is one that University Research Offices are increasingly keen to capitalise to their own advantage. Librarians, with the mandatory skills for managing information, are being tasked with developing online resources and making academic staff and student output available on repositories. The challenge facing libraries and special collections in South African (and African) Universities is the re-skilling of existing professional staff to undertake the new responsibilities that come revised job descriptions. University Librarians are hard pressed to find solutions.

Training is required at all levels, from management level to assistant level and covers a vast spectrum such as digital project management, metadata expertise for content and data management, digital media conversion, digital industry standards, best practices and guidelines. Many companies in South Africa do provide digital conversion services but not metadata creation services.

Training opportunities from commercial companies are few and far between, expensive, and often require extensive travel with long periods away from home. This is one of the roles that independent professionals can and are currently fulfilling.

¹¹http://stardata.nrf.ac.za/starweb/Carnegie/servlet.starweb?path=Carnegie/Fastlink.web&id=DIGIFL&pass=&search1=NORG%3DT*.U*&format=FastlinkReport (Accessed 27 August 2012).

¹² Liebetrau, *Managing Digital Collections*.

7. Independent Professionals

According to the website of The Association of Independent Information Professionals¹³ in the USA, “independent information professionals use high-level skills in finding, managing, applying, and communicating information, and they pursue their calling with an entrepreneurial spirit” ...and they ...”use the skills of librarians, private investigators, database searchers, market researchers, competitive intelligence researchers, writers, indexers, and other professions in their work.” This correlates with the observations of Wamundila et al.,¹⁴ who view knowledge and content management specialists as a special breed of information professionals. Furthermore these “super” information professionals manage knowledge that helps in reducing organisational operational costs, improves income flow and operational performance.

In the context of this paper the term Independent Professionals is used to include independent information *and* information technology professionals.

In South Africa, independent professionals are providing much needed skills and services in professional training and the application of innovative technology to support cultural heritage, library and archival digital resource requirements. They are plugging a gaping hole by providing an important body of knowledge and expertise, coupled with local understanding. They provide a level of service provision that is not otherwise available in the country, especially in the rapidly evolving technology environment where innovative solutions are required to suit an African context.

Independents are well placed to provide training that is at the cutting edge, but relevant, customised and suited to the level and understanding of trainees. Training for ongoing professional development is currently the most pressing segment required to hasten the rate of digital initiative development. This encompasses all aspects of the *new roles = new job descriptions = new skills* scenario accompanied by change management and leadership development requirements to support a solid foundation for this transformation. The Centre for African Library Leadership (CALL)¹⁵ at the University of Pretoria was developed to address the need for library leadership training and skills development of library managers in a formal programme approach. However, it is not only managers that benefit from leadership development.

Two case studies are discussed here as illustrative examples of the roles played by information professionals.

7.1 Mentoring Programme

A Mentor Programme has been designed and implemented to assist libraries expand their services and tackle new challenges in building digital resources. After adopting many training approaches it became evident that short courses tended to be unsatisfactory for many trainees as they didn’t have the understanding of all the “parts” fitting together. Frequently librarians would complain that they weren’t able to deal with the IT requirements or IT technicians would complain that they couldn’t understand librarian’s needs. Librarians and IT professionals have not traditionally in the past worked together but

¹³ <http://www.aiip.org/> (Accessed 27 August 2012).

¹⁴ Sitali Wamundila, Felesia Mulauzi, Naomy Mtanga, and Benson Njobvu, “Meeting the training needs for knowledge and content management specialists: case study of Southern African Universities” (paper presented at SCECSAL 2012 Conference, Nairobi, Kenya, 4-8 June 2012), p. 5.

¹⁵ <http://www.library.up.ac.za/carnegie/centre.htm> (Accessed 27 August 2012).

now need to work closely with each other now and it seems that it is not a good marriage. Thus the “whole (integrated) picture” was required to be experienced and understood, over a period of time, in a supportive, non-threatening environment. The mentor programme was developed to satisfy this need. A unique aspect of the programme is the specialised and individual approach tailored to each Mentee while satisfying the requirements of their own institution. This approach is recommended in findings in an investigation to establish the (then) current state of University libraries in Africa.¹⁶ Extensive hands-on training is accompanied by development of guidelines, best practices and quality assurance processes. Assisting organisations in documenting policies, guidelines and best practices are essential in saving time and money due to changes in management and staff turnover. No point in re-inventing the wheel. Documentation can be versioned as changes are introduced and adapted for various projects. They can be made available on an organisational library website and used as a template by others to make the process easier and more efficient.

The mentoring programme is particularly useful for the more geographically remote Universities in African countries where little or no specialised training facilities are available locally. This approach equips Mentees with complete confidence in dealing with project management, scoping, assessing equipment requirements, metadata creation, long-term preservation, and ensuring high quality deliverables at the end of the project. A train-the-trainer approach, where the trainee is tasked with training colleagues and staff back home after completing the Mentor programme, has proved economically viable for many small libraries.

7.2 The Ulwazi Community Memory Programme (<http://www.ulwazi.org/>)

Another innovative way that independent professionals are making a difference is in the use of mobile technologies to assist communities build indigenous knowledge resources.

Durban is a busy port city on the east coast of South Africa. In the early part of the twentieth century Durban’s development as a colonial city, was clearly reflected in styles of architecture, in dress codes, in modes of transport and others. A hundred years later, and the city has developed into a thriving multicultural commercial and industrial African city, with a population of around 3 million inhabitants. The local museums and archives in Durban hold rich resources about the local history but it is the Zulu heritage which is the least documented and urban migration threatens the loss of oral histories traditionally passed down through families. The eThekweni Municipality of Durban has an infrastructure of around 89 public libraries. The Ulwazi Project was developed, with funding assistance from the Goethe Institute, as part of the Public Library’s mandate to establish online indigenous resources as an integral part of services to the communities they serve. It comprises a portal, wiki and blog where indigenous knowledge from the local Durban communities can be submitted, uploaded and preserved using Web 2.0 technologies. “The utilisation of the combination of open-source and social media applications for archival and heritage purposes makes this project unique in South Africa.”¹⁷

¹⁶ Diana Rosenberg, *Towards the digital library: findings of an investigation to establish the current status of university libraries in Africa* (Oxford: International Network for the Availability of Scientific Publications (INASP), 2005), 17.

¹⁷ Elizabeth Greyling and Niall McNulty, “How to build an Indigenous Digital Library through Community Participation: The case of the Ulwazi Programme” (paper presented at SCECSAL 2012 Conference, Nairobi, Kenya, 4-8 June 2012), p. 9.

Following on the success of the original Project, the Ulwazi Programme sought ways to develop an innovative model to capitalise on the phenomenal growth of mobile technology, specifically those with browser technologies. It is estimated that cell phone usage in Africa is now close to 70%.¹⁸ To make the project more relevant to the local community, the content on the website was made accessible to users through their cell phones. Then the concept was extended to enable the collection of indigenous knowledge material using mobile phones. “The initiative showed that cell phones could be used productively as a tool in the exchange of indigenous knowledge...”¹⁹ This enhances the archival knowledge experience for members of local communities. The eThekweni Municipal Library has assumed responsibility for the long-term curation of the data.

8. Preservation of Digitized Resources

“In sub-Saharan Africa, the volume of digital information media produced is comparatively small, but preservation challenges are particularly acute. If digital media cannot be preserved, part of the Africa’s heritage is being lost.”²⁰

Despite setbacks and challenges, there are important projects digitising heritage resources and providing access. Some of these projects have stalled or been parked for lack of funds to support long-term preservation. Very few are actively focused on preservation issues in a sustainable way. One could argue that digitization in itself is a form of preservation—by making a digitized surrogate, the original physical form is less handled and better preserved. Also by making resources available you are (in a way) preserving heritage by keeping it alive. But the data also needs to be kept alive...! Cohesive efforts will need to be implemented as more resources are digitized. The lesson learned from DISA is that long-term curation and preservation is not an add-on luxury but requires attention from the beginning of the project.

9. Conclusion

A measure of the contribution that independent professionals are making is difficult to assess. In the short term, tangible success could be measured by increased digitized content, websites, available guidelines and policy documents. Referrals and requests from satisfied clients are always a good measure.

Currently, independent professionals work largely in a loose networked environment, respecting each other’s skills and professional ethics. Three of the final recommendations from the INASP investigation into digital library needs in Africa²¹ were:

- Best practice in user education for the digital environment should be summarised and disseminated to ensure efficient use of digital library services.
- Working with partners, develop and support continuing education and training programmes for librarians using a variety of approaches and methodologies.

¹⁸ Ibid.

¹⁹ Ibid., p. 11.

²⁰ Peter Johan Lor, “Preserving Digital African Resources: Is there a role for repository libraries,” *Library Management* 26, no. 1/2 (2005): 63-72, p. 63.

²¹ Rosenberg, *Towards the digital library*, p. 28.

- Support consortia to build strong networks and expertise within their countries/regions, so enabling them to take on wider coordination and advisory roles and to foster collaboration among libraries involved.

It is possible that a more formal Association could benefit the heritage sector by playing a more proactive role in developing tools, promoting advocacy and fostering new partnerships, cross-institutional collaborations and mentoring initiatives. This would require a subscription model to sustain and provide for staff and operational costs. In the interim it can be said that independent professionals can, and are, making a positive, if modest, contribution to the state of the art of preserving our culture.

References

- Dunlop, Janine and Lesley Hart. "Digitisation projects at the University of Cape Town Libraries." *Innovation* 30 (June 2005): 32-42.
- Greyling, Elizabeth and Niall McNulty. "How to build an Indigenous Digital Library through Community Participation: the case of the Ulwazi Programme." Paper presented at SCECSAL 2012 Conference, Nairobi, Kenya, 4– 8 June 2012.
- Liebetrau, Pat, ed. *Managing Digital Collections: a collaborative initiative on the South African Framework*. Pretoria, South Africa: National Research Foundation, 2010.
- Lor, Peter Johan. "Preserving Digital African Resources: is there a role for repository libraries?" *Library Management* 26, no. 1/2 (2005): 63-72. DOI 10.1108/01435120510572888.
- Lor, P. J. "Digital libraries and archiving knowledge: some critical questions." *South African Journal of Libraries and Information Science* 74, no. 2 (2008): 117-128.
- May, Julian. *Digital and other poverties: Exploring the connection in four East African countries*. Durban: University of KwaZulu-Natal, 2010.
- Rosenberg, Diana. *Towards the digital library: findings of an investigation to establish the current status of university libraries in Africa*. Oxford: International Network for the Availability of Scientific Publications (INASP), 2005.
- South Africa, Department of Arts and Culture. "National Policy on the Digitisation of Heritage Resources." Final Draft for Public review, 2010.
- Wamundila, Sitali, Felesia Mulauzi, Naomy Mtanga, and Benson Njobvu. "Meeting the training needs for knowledge and content management specialists: case study of Southern African Universities." Paper presented at SCECSAL 2012 Conference, Nairobi, Kenya, 4– 8 June 2012.

Is A New Legal Framework Required for
Digital Preservation or Will Policy Do?

Is a New Legal Framework Required for Digital Preservation or Will Policy Do?

Building a Legal Framework to Facilitate Long-term Preservation of Digital Heritage: A Canadian Perspective

Tony Sheppard

Professor, Faculty of Law, The University of British Columbia, Canada, sheppard@law.ubc.ca

Abstract

What is “law” and how can law reform facilitate digitalization and preservation of cultural heritage? The current UNESCO Charter on the Preservation of Digital Heritage contains an excellent start, but would benefit from defining its terms and broadening its scope to add issues that States should consider in developing their laws. These issues start with State constitutions, and include tax reform, copyright law reform, protection and preservation of the cultures of indigenous peoples, and the practical problems of law enforcement in the digital era. Some Canadian solutions are discussed in this paper. It is also suggested that the role of UNESCO as contained in the current Charter could be expanded to include the drafting of model provisions with commentary to assist States in improving their domestic laws.

Author

Tony is a Professor in the Faculty of Law, The University of British Columbia, and a retired member of the Law Society of British Columbia. His research interests include the law of evidence and admissibility of digital records. He currently teaches courses in taxation law, law of equity and real property law in the Faculty and supervises graduate students. He is actively involved in law reform and interdisciplinary projects in collaboration with members of UBC’s School of Library, Archival and Information Studies (SLAIS) and with many others around the world.

1. Introduction

The provocative topic of this panel, “Is a new legal framework required for digital preservation or will policy do?” might be misunderstood as it appears to offer a false dichotomy between law and policy as alternatives, and that it would be open to a society governed by the rule of law, to bypass the formalities of law by adopting the expedient of policy. In a state where the rule of law prevails, it is impossible to divorce policy from law, however. Policy is subject to the rule of law through judicial scrutiny. Governmental policy must be founded on legislation. Under the rule of law, Parliament must enact the legislative basis for the iteration of policy. On this legislative basis the Executive can develop policy, which would take the form of delegated legislation, known as “regulations.” The rule of law only requires that the bare bones of Parliamentary intent to authorize the Executive to make regulations appear in the statute, leaving plenty of scope for policy development in regulations. The rule of law requires that regulations must be subject to judicial review to ensure that they do not exceed the underlying legislation.¹ The *United Nations Millennium Declaration* articulates the supremacy of the rule of law and connects it with digitization and cultural heritage.²

The author gratefully acknowledges the many helpful comments and suggestions offered by Ken Cavalier, Ph.D., LL.M., LL.B., and B.A. on a draft of this paper.

¹ *Catalyst Paper Corp v North Cowichan (District)*, 2012 SCC 2, [2012] 1 SCR 5 at paras 12-15.

² UNGAOR, 55 Sess, UN Doc 55/2 (2000), Articles 20 [digitization], 24 [rule of law] and 25 [cultural heritage].

Law features prominently in the UNESCO, *Charter on the Preservation of Digital Heritage*.³ First of all, the *Charter* defines the scope of digital heritage to include “legal” heritage materials.⁴ Second, the *Charter* identifies “the lack of supportive legislation” as a factor contributing to the endangerment of digital heritage worldwide.⁵ Third, the *Charter* urges States to adopt legal measures safeguarding digital heritage.⁶ Fourth, and giving rise to the topic of this article, the *Charter* recommends that States enact appropriate legal frameworks to secure the protection of their digital heritage.⁷ The *Charter* states that a legal framework for the protection and preservation of digital heritage should provide for legal deposit and for reasonable access by the public to the deposited heritage material.⁸ States are also encouraged to develop another legal framework for ensuring “authenticity” of digital heritage and to prevent its “manipulation or intentional alteration.”⁹

This paper offers some suggestions whereby the law might cease to be part of the problem, as is alleged by the *Charter*, Article 3, and become part of the solution in eliminating or reducing impediments to the preservation of digital heritage. Protection of cultural heritage is not only an important Canadian constitutional value, but also a worthwhile endeavor in its own right.¹⁰

What reforms might be included in a State’s effort at building a comprehensive series of legal frameworks to facilitate long-term preservation of digital heritage? A revised *Charter* might define “legal frameworks” to inform States of the broad scope of issues to be addressed, including constitutional, fiscal and taxation, and copyright impediments to the preservation and protection of cultural heritage. The perilous future of indigenous cultural heritage invokes the honour of the Crown as protector of human rights and cultural diversity in relation to Canada’s First Nations, Métis and Inuit peoples.

The *Charter* raises another issue: what is the most constructive role for UNESCO to play in regard to the legal aspects of preserving and protecting digital heritage? Article 12 of the *Charter*, which defines “[t]he role of UNESCO” only refers to proposing standard “**legal**” and technical guidelines, to support the preservation of the digital heritage.”¹¹ A revised and amended *Charter* might propose a broader role for UNESCO in building model legal frameworks to overcome specific legal impediments, which it would recommend for adoption by State legislatures.

³ UNESCO, *Charter on the Preservation of Digital Heritage*, online: http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html (15 October 2003) [*Charter*].

⁴ *Ibid.*, Article 1, which identifies “digital heritage” as follows: “The digital heritage consists of unique resources of human knowledge and expression. It embraces cultural, educational, scientific and administrative resources, as well as technical, legal, medical and other kinds of information created digitally, or converted into digital form from existing analogue resources. Where resources are “born digital”, there is no other format but the digital object. Digital materials include texts, databases, still and moving images, audio, graphics, software and web pages, among a wide and growing range of formats. They are frequently ephemeral, and require purposeful production, maintenance and management to be retained.”

⁵ *Ibid.*, Article 3.

⁶ *Ibid.*, Article 4.

⁷ *Ibid.*, Article 8. For an explanation of “legal deposit,” please see Library and Archives Canada, *Legal Deposit*, online: <http://www.collectionscanada.gc.ca/legal-deposit/index-e.html>.

⁸ *Ibid.*

⁹ *Ibid.*

¹⁰ See *Windsor (City) v Paciorka Leaseholds Limited*, 2012 ONCA 431 at para 27 (cultural heritage resources provide economic, environmental and social benefits).

¹¹ *Charter*, *supra* note 3, Article 12 [emphasis added].

2. The Nature and Sources of Law

What is law? International bodies' aspirational statements about the role of cultural heritage in economic and societal development are calls for immediate action, but they do not have the force of law.¹² Unlike laws, they do not require obedience or penalize disobedience.

International law comprises international conventions, treaties and declarations, but domestic enforceability requires legislation that has been enacted locally by a competent legislature rather than international instruments that have not attained implementation by local legislation. Implementation of international law into the domestic law of Canada generally requires federal, and sometimes provincial, legislation. Implementation of international law can occur in either of two ways: either by enactment of new legislation bringing the treaty into effect or by the presence of existing federal and provincial legislation that is consistent with the new treaty.¹³

Upon ratification of a treaty, Canada becomes bound by it and cannot enact legislation contradicting it.

Even without implementation by domestic legislation, expressions of international values and aspirations are influential with judges and may influence domestic case law, particularly judicial interpretation of domestic legislation.¹⁴

3. Treaties

On the topic of the preservation of digital cultural heritage, the most significant treaty ratified by Canada is the UNESCO *Convention Concerning the Protection of the World Cultural and Natural Heritage*.¹⁵

The Treaty¹⁶ refers to law only twice. First, in Article 5.4, the Treaty¹⁷ refers to law as follows:

Article 5

To ensure that effective and active measures are taken for the protection, conservation and presentation of cultural and natural heritage situated on its territory, each State Party

¹² E.g., Commonwealth Foundation (2008), *Putting Culture First* (London: The Commonwealth Foundation, 2008), online: <http://www.commonwealthfoundation.com/LinkClick.aspx?fileticket=16LU0GdSEto%3D&tabid=247>; Commonwealth Foundation (2010), *Commonwealth Statement on Culture and Development: Prepared by the Commonwealth Group on Culture and Development* (London: Commonwealth Foundation, 2010), online: <http://www.commonwealthfoundation.com/LinkClick.aspx?fileticket=n4AYwvbYDEI%3D&tabid=167>.

¹³ Parliament of Canada, Library of Parliament Research Publications, *Canada's Approach to the Treaty-making Process*, by Laura Barnett (Background Paper No. PRB 08-45-E, 24 November 2008), online: <http://www.parl.gc.ca/Content/LOP/ResearchPublications/prb0845-e.htm>; see also *Society of Composers, Authors and Music Publishers of Canada*, 2004 SCC 45, [2004] 2 SCR 427.

¹⁴ *Baker v Canada (Minister of Citizenship and Immigration)*, [1999] 2 SCR 817, 174 DLR (4th) 193, at paras 69-72; *R v Sharpe*, [2001] 1 SCR 45, 2001 SCC 2 at para 175; *Németh v Canada (Justice)*, 2010 SCC 56, [2010] 3 SCR 281, at paras 34-35. The Canadian Parliament is presumed not to legislate in breach of a treaty, the comity of nations and the principles of international law: see *Daniels v. White*, [1968] SCR 517, at p 541.]

¹⁵ UNESCO *Convention Concerning the Protection of the World Cultural and Natural Heritage*, 23 July 1976, 1037 UNTS 151, CanTS 1976/44 (entered into force 17 December 1975) [Treaty].

¹⁶ Ibid.

¹⁷ Ibid.

to this Convention shall endeavor, in so far as possible, and as appropriate for each country:...

4. to take the appropriate **legal**, scientific, technical, administrative and financial measures necessary for the identification, protection, conservation, presentation and rehabilitation of this heritage; [emphasis added]

What does UNESCO mean by the phrase, the “appropriate legal measures?”¹⁸ Giving specificity to the phrase, UNESCO refers to a State’s implementation of treaties as legislation protecting treaty rights within the State’s “legal framework,”¹⁹ but does not define the term. The *Charter* also refers to but does not define “legal framework.”²⁰

At its most fundamental level, a State’s legal framework is based on protection of treaty rights by its “constitution.” UNESCO also includes within a State’s legal framework, protection of treaty rights by “a basic legislative text or by any other national provision.”²¹ Incorporation of the UNESCO treaty into “national legislation” also qualifies as a building block within a State’s legal framework.²² Mixing law and administration, UNESCO includes a State’s conferring of authority in relation to treaty rights on a competent authority in the jurisdiction as part of its legal framework.²³

The second reference to law is contained in Article 34 of the Treaty,²⁴ which recognizes that States, such as Canada, having a federal or non-unitary constitutional system divide their legislative powers among legislative bodies and that complete implementation of treaty obligations may also be divisible between legislatures. For example, the Canadian federation divides legislative competence between a Federal Parliament in Ottawa, and provincial and territorial legislative assemblies located within each province or territory. Each of these legislative bodies is sovereign within the legislative powers assigned to it by the Canadian constitution.

Article 34 requires a federal government to implement the convention as far as it is able within its constitutional powers and to inform the other provincial or territorial governments of the treaty obligations within their legislative competence with a recommendation for adoption by their legislatures.

Because the Internet transcends territorial legislative boundaries, it can be important for such purposes as copyright or taxation to identify which level of government can legislate extra-territorially. In Canada, only the federal government can do so, the provinces and territories can impose their legislation only within their respective borders.²⁵

Canada has been a party to other significant treaties on cultural property, however, such as the following:

¹⁸ Ibid.

¹⁹ UNESCO, Executive Board, “Framework Guidelines,” 177EX/Decision 35 II, 177th Sess., 01 Oct 2007 at para 1.(b); online: http://portal.unesco.org/en/ev.php-URL_ID=41897&URL_DO=DO_PRINTPAGE&URL_SECTION=201.html [*Framework Guidelines*].

²⁰ *Charter*, *supra* note 3.

²¹ *Framework Guidelines*, *supra* note 19.

²² Ibid.

²³ Ibid.

²⁴ Treaty, *supra* note 15.

²⁵ *Society of Composers, Authors and Music Publishers of Canada v Canadian Association of Internet Providers*, 2004 SCC 45, [2004] 2 SCR 427 [*Society of Composers, Authors and Music Publishers*].

Title	Date of Signature or Accession	Treaty Series Reference	Other information
<i>Constitution of the United Nations Educational, Scientific and Cultural Organization</i>	16 November 1945	4 UNTS 75; CanTS 1945/18	(entered into force 04 November 1946)
<i>Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property</i>	28 March 1978	CanTS 1978/33	(entered into force 24 April 1972; implemented by <i>Cultural Property Export and Import Act</i> , RSC 1985, c C-51)
<i>Convention for the Protection of Cultural Property in the Event of Armed Conflict</i>	11 December 1998	249 UNTS 240; Can TS 1999/52	(entered into force 07 August 1956)
<i>Protocol for the Protection of Cultural Property in the Event of Armed Conflict</i>	29 November 2005	249 UNTS 358; CanTS 2006/26	(entered into force 07 August 1956) (referred to as “The First Protocol”)
<i>Second Protocol to the Hague Convention of 1954 for the Protection of Cultural Property in the Event of Armed Conflict</i>	29 November 2005	CanTS 2006/27	(entered into force 09 March 2004)
<i>Convention on the Protection and Promotion of the Diversity of Cultural Expressions</i>	28 November 2005	CanTS 2007/8	(entered into force 18 March 2007)

4. Building the Legal Framework

4.1. The Constitution

Building a legal infrastructure for digitalizing and preserving cultural heritage can start by laying a strong foundation. The foundation of a country’s laws and society is its constitution.²⁶ Therefore, in building a legal framework for the preservation of digital cultural heritage within a State, the starting-point should be consideration of whether or not to reform its constitution. Canada’s constitution comprises statutes and related principles. An important part of the Canadian constitution is a statute entitled, *The Canadian Charter of Rights and Freedoms*,²⁷ which in turn is based on foundational constitutional principles, such as the rule of law²⁸ and respect for the (minority) French language and culture as expressed in sections 16-24, and for multiculturalism as expressed in section 27. Section 27 of the *Canadian Charter of Rights and Freedoms* states as follows:

²⁶ *Citizen’s and The Queen Insurance* (1880), 4 SCR 215; *Re Manitoba Language Rights*, [1985] 1 SCR 721; *Quebec (Attorney General) v Laroche*, 2002 SCC 72, [2002] SCR 708; *British Columbia v Imperial Tobacco Canada Ltd*, 2005 SCC 49, [2005] 2 SCR 473.

²⁷ *Canadian Charter of Rights and Freedoms*, Part 1 of the *Constitution Act*, 1982, being Schedule B to the *Canada Act 1982* (UK), 1982 c 11.

²⁸ *Ibid.*, Preamble to the *Canadian Charter of Rights and Freedoms*.

Multicultural heritage

27. This Charter shall be interpreted in a manner consistent with the preservation and enhancement of the multicultural heritage of Canadians.

Section 27 recognizes the fundamental principle that the heritages of the French, First Nations, Métis and Inuit peoples and other cultural groups are worthy of preservation and enhancement.²⁹ The revised and amended UNESCO *Charter* should define “legal framework” and include a State’s constitution in the definition.

4.2 Legislation: the preferable method of building a legal framework

Domestically, States and their subdivisions are ruled by constitutional laws, legislation and case law. Case law is an important source of domestic law, but court decisions can only effect incremental change in the law through individual cases and cannot establish a comprehensive legal framework for change. For a comprehensive legal solution to a problem such as the preservation of cultural heritage by digital technology, legislation following a review of the legal consequences is the preferred method of achieving meaningful reform.³⁰ The rule of law is satisfied by a basic outline in legislation, with authority to formulate policy delegated to the Executive through the making of regulations pursuant to the statute. This arrangement facilitates responsive rule-making on a rapidly evolving topic such as digital preservation.

Domestic legislation, within limits permitted by the local constitutional law of the jurisdiction, would be necessary to provide a valid and enforceable legal framework for facilitating the long-term preservation of digital heritage. In framing such legislation, a legislature can incorporate concepts developed in international law, foreign law, or case law. The drafters of legislation can borrow legal terms and concepts from anywhere or create their own, as permitted within constitutional limits. Also, proposed legislation should be consistent with other established legislative provisions and case law of the local jurisdiction.

Harmonization of the domestic laws of different jurisdictions is also desirable, to improve efficiency and certainty, in dealing with a common problem such as the preservation of digital heritage. A possible route would be for each jurisdiction, separately and independently, to develop its own domestic legislation dealing with the preservation of its digital heritage. So far, results of this approach are mixed: see the UK’s attempted *Digital Heritage Bill 2010* and Quebec’s *An act respecting the governance and management of information resources of public bodies and government enterprises*.³¹ Turkey’s legislative attempt to protect its cultural heritage, *The Cultural and Natural Heritage Protection Act 2863*, is well-intentioned but has been criticized as unfair to property rights.³²

²⁹ See *R v Monteith*, (1991) 132 NBR (2d) 203, 5 CR (4th) 241 (QB).

³⁰ *Watkins v Olafson*, [1989] 2 SCR 50, 61 DLR (4th) 577; *R v Salituro*, [1991] 3 SCR 654, 68 CCC (3d) 289; *WIC Radio Ltd v Simpson*, 2008 SCC 40, [2008] 2 SCR 420; *Myers v Director of Public Prosecutions*, [1965] AC 1001 (HL).

³¹ An act respecting the governance and management of information resources of public bodies and government enterprises, RSQ, c G-1.-3.

³² Compare Sevil Yildiz, “The Model of Turkey in Legal Protection of Cultural Heritage” International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVIII, Part 5, Commission V Symposium, Newcastle Upon Tyne, UK, 2010, p 627. Online: <http://www.isprs.org/proceedings/XXXVIII/part5/papers/119.pdf>; with Nida Celik and Bayrum Uzun, “Cultural Heritage versus Property Rights” TS09C-Surveying and Cultural Heritage II, 5686; FIG Working Week 2012; Knowing to manage the territory, protect the environment, evaluate the cultural heritage; Rome, Italy, 6-10 May 2012, online: http://www.fig.net/pub/fig2012/papers/ts09c/TS09C_celik_uzun_5686.pdf.

From an international perspective, if each State develops its legislation on its own, the result would be a patchwork of domestic laws at best. Collective effort under the auspices of an international body such as UNESCO might be more effective to harmonize domestic laws. Since 2005, UNESCO has been developing a *Database of National Cultural Heritage Laws*.³³ Proposing model legislation would be a logical next-step for UNESCO to take.

The revised and amended *Charter on the Preservation of Digital Heritage*³⁴ might articulate a broader role for UNESCO than the current Article 12 provides for.

Is it worth considering a role for UNESCO in the development of a model enactment with commentary at the global level? The model provisions would draw upon concepts from local precedents and from the work of others. UNESCO might recommend the model enactment as suitable for adoption by domestic legislatures, with scope for variations as might be considered necessary in each jurisdiction and as described in the commentary to the legislation. The overarching objective of the project would be international harmonization of domestic laws as much as possible. The proposed commentary to the model provisions could contain an informative comparison of the diverse domestic laws on such common issues as authentication of digital records, intellectual property rights, protection of privacy, e-discovery, existing legal duties to preserve digital information and spoliation. Would this be an appropriate and attainable task for UNESCO?

4.3. Suggested topics for additional legal frameworks

For greater certainty, a revised and amended *Charter on the Preservation of Digital Heritage*³⁵ might indicate which topics are significant legal frameworks to improve the preservation and protection of a State's digital heritage. As discussed in the pages that follow, tax reform could create a fiscal framework for donations and other resources to be applied to digital heritage. A framework for copyright law could facilitate digital heritage, too. Finally, the endangered cultural heritage of indigenous peoples should be contained in a legislative framework that is sensitive to the issues.

4.3.1. Funding: tax law reform

The UNESCO *Convention Concerning the Protection of World Cultural and Natural Heritage*,³⁶ Article 5.4, requires States to take appropriate “**financial measures** necessary for the identification, protection, conservation, presentation and rehabilitation of [natural and cultural] heritage.” The *Charter on the Preservation of Digital Heritage*,³⁷ in Article 3, says: “Digital evolution has been too rapid and costly for government and institutions to develop timely and informed preservation strategies.” If the costs of digital protection and preservation of cultural heritage exceed public resources, recourse to private resources is necessary. The *Charter* does not spell this option out clearly: Article 4 merely states: “Member States will benefit by encouraging **legal, economic** and technical measures to safeguard the heritage.”³⁸

³³ UNESCO *Database of National Cultural Heritage Laws*, online: <http://www.unesco.org/culture/natlaws>.

³⁴ *Charter*, *supra* note 3.

³⁵ *Charter*, *supra* note 3.

³⁶ Treaty, *supra* note 15.

³⁷ *Charter*, *supra* note 3.

³⁸ *Ibid.*, (emphasis added).

A revised *Charter on the Preservation of Digital Heritage* could make explicit reference to the need for a legal framework dealing with fiscal incentives to encourage private funding and allocation of other private resources to the protection and preservation of digital cultural heritage. Legislation is necessary to authorize the imposition of taxes, and incentives are commonly introduced to encourage heritage preservation, such as heritage buildings. Projects for the digitalization and preservation of cultural heritage require funding to pay for supplies, equipment and personnel. Though the cost of digital equipment might be falling, other costs rise continually. Government grants can be made available, but current economic realities constrain public funding, necessitating greater recourse to private sources of capital for funding. In a market economy, cultural heritage may not compete favorably for funding against other less risky and more lucrative endeavors. Canadian law currently offers a wide range of exemptions from legal obstacles and provides numerous financial incentives to encourage investing in cultural heritage, because such heritage benefits society.

Canadian courts recognize that technological advances, on balance, are beneficial to society and that the law should encourage their lawful usage.³⁹ Digitalization of cultural heritage increases its availability to the public over the Internet. The resulting efficiencies, cost reductions and expanding accessibility benefit the public. Canadian law regards philanthropic or not-for-profit funding of digitalization and preservation of cultural heritage as charitable. Donations to charitable activities qualify for tax credits. More generous tax incentives could increase the allocation of private resources to these activities.

A revised *Charter* could encourage States to develop a fiscal framework for the protection and preservation of digital cultural heritage through new tax incentives encouraging private investment. The framework should encourage legitimate, good faith preservation, but penalize attempted abuses of the tax incentives.

4.3.2. Intellectual property rights: copyright law reform

Fear of the possible consequences of a copyright violation should not deter digital preservation of imperiled material by public institutions to prevent irretrievable loss of cultural property. Canada proposes to reduce this obstacle to the preservation of endangered material by exempting it from claims of copyright infringement. When implemented, Canada's *Copyright Modernization Bill*⁴⁰ will add an expanded paragraph 30.1(1)© to the *Copyright Act*,⁴¹ stating as follows:

30.1(1) It is not an infringement of copyright for a library, archive or museum or a person acting under the authority of a library, archive or museum to make, for the maintenance or management of its permanent collection or the permanent collection of another library, archive or museum, a copy of a work or other subject-matter, whether published or unpublished, in its permanent collection

...

(c) in an alternative format if the library, archive or museum or a person acting under the authority of the library, archive or museum considers that the original is currently in a format that is obsolete or is becoming obsolete, or that the technology required to use the original is unavailable or is becoming unavailable;

³⁹ *BMG Canada Inc v John Doe*, 2005 FCA 193, [2005] 4 FCR 81 at para 41; *Voltage Pictures LLC v Jane Doe*, 2011 FC 1024 at para 14; *R v Schroeder*, 2012 ABPC 241 at para 24.

⁴⁰ Canada, Bill C-11, *The Copyright Modernization Act*, 1st Sess, 41st Parl, 2011 (assented to 29 July 2012).

⁴¹ *Copyright Act*, RSC 1985, c C-42.

The Bill unfortunately permits copyright holders to over-ride this otherwise laudable provision, among others, with the installation of inviolable digital locks. Despite criticism⁴² of digital locks, the Canadian government persevered with its provisions upholding the primacy of locks.

The problem of “orphan works” obstructs efforts in various States to digitalize vast amounts of cultural heritage and threatens loss of these works to posterity, unless legislation is enacted to resolve the dilemma. When copyright protects a work from use without consent of the current owner, and the owner of the copyright cannot be identified or located to give consent, the work is said to be an “orphan.” In the absence of legislation, usage of the orphan work is frozen.

Section 77 of the Canadian *Copyright Act*⁴³ removes this stumbling block for orphan works covered by Canadian copyright. Basically, the person who wishes to use a work that is subject to Canadian copyright held by someone whose identity or whereabouts cannot be ascertained (called the “unlocatable copyright owner”), may apply to the Copyright Board of Canada for permission to copy the work. If the application is successful, the applicant receives a licence from the Board permitting the proposed use of the work, which prevents the unlocatable owner from suing the applicant for breach of copyright. The Board has a discretion to grant such permission by way of a non-exclusive licence upon being satisfied that the work is subject to a current Canadian copyright and the applicant has made reasonable but unsuccessful efforts to identify and locate the copyright owner. The Board can impose terms and conditions on granting the licence that require the user to pay a (usually) nominal royalty, which the unlocatable owner may claim anytime up to five years after expiry of the licence. In practice, unlocatable owners rarely come forward to claim their modest royalties. The Canadian regime for authorizing usage of orphan works has gained general acceptance in this country, but its adoption by other jurisdictions remains a live issue. The United Kingdom and the European Union are currently in the process of enacting controversial proposals.⁴⁴

Canada’s orphan work provision, section 77 of the *Copyright Act*,⁴⁵ provides as follows:

Section 77 (1) Where on application to the Board by person who wishes to obtain a licence to use

- (a) a published work,
- (b) a fixation of a performance,
- (c) a published sound recording, or
- (d) a fixation of a communication signal

in which a copyright subsists, the Board is satisfied that the applicant has made reasonable efforts to locate the owner of the copyright and that the owner cannot be located, the Board may issue to the applicant a licence...

(2) A licence issued under subsection (1) is non-exclusive and is subject to such terms and conditions as the Board may establish.

(3) The owner of a copyright may, not later than five years after the expiration of a licence issued pursuant to subsection (1) in respect of the copyright, collect the royalties fixed in the licence or, in default of their payment, commence an action to recover them in a court of competent jurisdiction.

⁴² Canadian Library Association, *Call to Action on Copyright: Bill C-11* (January 2012), online: www.CLA.ca.

⁴³ Copyright Act, *supra* note 41.

⁴⁴ UK, Bill 61, *Enterprise and Regulatory Reform Bill*, 2012-2013 sess., 2012; s. 59; EU Directive, *Certain Permitted Uses of Orphan Works*, passed by the European Parliament, 13 September 2012.

⁴⁵ Copyright Act, *supra* note 41.

The revised and amended *Charter on the Preservation of Digital Heritage*⁴⁶ should identify the need for a legislative framework to deal with copyright, and list these two issues, imperiled and orphan works, as primary issues for resolution in the framework. Is there a role for UNESCO in developing optimal legislative solutions to the thorny problems of protecting digital works from loss due to changing technology and overcoming copyright obstacles concerning orphan works?

4.3.3. Digitizing and Preserving the Cultural Heritage of Indigenous Peoples

The United Nations *Declaration on the Rights of Indigenous Peoples*⁴⁷ contains many references to States' responsibilities in protecting and preserving the cultural heritage of indigenous peoples. Article 31 of the *Declaration*⁴⁸ is especially pertinent, stating as follows:

Article 31

1. Indigenous peoples have the right to maintain, control, protect and develop their cultural heritage, traditional knowledge and traditional cultural expressions, as well as the manifestations of their sciences, technologies and cultures, including human and genetic resources, seeds, medicines, knowledge of the properties of fauna and flora, oral traditions, literatures, designs, sports and traditional games and visual and performing arts. They also have the right to maintain, control, protect and develop their intellectual property over such cultural heritage, traditional knowledge, and traditional cultural expressions.

2. In conjunction with indigenous peoples, States shall take effective measures to recognize and protect the exercise of these rights.

After initially opposing the *Declaration on Indigenous Peoples*, Canada reversed its position and approved of it, as an aspirational statement.⁴⁹ The revised and amended *Charter on the Preservation of Digital Heritage* might make specific reference to development of a legal framework relating to the digital preservation of the cultural heritage of indigenous peoples, as especially significant. Consistently with Article 31, Canadian courts recognize the cultural significance of Aboriginal rights over their heritage.⁵⁰

4.3.4. Enforcement by States of Treaty Obligations and Other Responsibilities in the Digital Era

In recent decades the volume of illicit trade in cultural objects has grown spectacularly and beyond traditional legal control. To counteract this abuse, UNESCO recommends⁵¹ that States should

⁴⁶ *Charter*, *supra* note 3.

⁴⁷ Declaration on the Rights of Indigenous Peoples, GA Res 295, UNGAOR, 61st Sess, Supp No 49, UNDOC A/RES/61/295 (2007) [Declaration on Indigenous Peoples].

⁴⁸ *Ibid.*

⁴⁹ Aboriginal Affairs and Northern Development Canada, *Backgrounder: Canada's Endorsement of the United Nations Declaration on the Rights of Indigenous Peoples*, (Ottawa: AAND, 2010), online: <http://www.aadnc-aandc.gc.ca/eng/1292353979814/1292354016174>.

⁵⁰ *William v British Columbia*, 2012 BCCA 285 at para 171.

⁵¹ UNESCO, *Illicit Traffic of Cultural Property*, online: <http://www.unesco.org/new/en/culture/themes/movable-heritage-and-museums/illicit-traffic-of-cultural-property/>.

immediately ratify and implement the 1970 *UNESCO Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property*.⁵² The prevalence of E-commerce poses challenges to the regulation of international transactions in cultural heritage property. Transactions by email or through websites transcend geographical borders and do not occur in places known to traditional law.

Following the definitions in international law, Canadian export/import controls on cultural heritage property apply only to certain types of tangible movable cultural property.⁵³ E-commerce threatens to put electronic trade in cultural heritage property beyond the practical reach of domestic law. The Internet “exists” in cyberspace, transcending national boundaries, yet laws are usually territorial rather than international. This geographical limitation challenges the effectiveness of attempting to regulate transactions over the Internet by domestic law.⁵⁴ UNESCO seeks to avoid such impracticalities by taking a broader approach, to the effect that “all countries should attempt to respond to the illicit trade in cultural objects via the Internet by taking appropriate measures,” and offers practical suggestions for trade barriers to impede illicit Internet dealings.⁵⁵ Domestic laws can help to regulate the sale and export of culturally significant materials by requiring export permits and conferring pre-emptive rights of first refusal on local purchasers, prior to export. Should a revised and amended *Charter on the Preservation of Digital Heritage*⁵⁶ offer guidance to States on a model legal framework for enforcement?

5. Conclusion

A global approach to building domestic legal frameworks for the preservation of digital heritage is attainable with the active involvement of UNESCO and preferable to individual States proceeding on their own. UNESCO should review its instruments and increase its activities to facilitate this most worthwhile but complex of human endeavors.

⁵² UNESCO *Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property*, 28 March 1978, 823 UNTS 231, CanTS 1978/33, (entered into force 24 April 1972).

⁵³ Canadian Heritage- Movable Cultural Program, *A Guide to Exporting Cultural Property from Canada*, online: <http://www.pch.gc.ca/pgm/bcm-mcp/frm/guide-eng.cfm>.

⁵⁴ *Society of Composers, Authors and Music Publishers*, *supra* note 25.

⁵⁵ UNESCO, *Basic Actions concerning Cultural Objects being offered for Sale over the Internet*, online: <http://portal.unesco.org/culture/fr/files/21559/11836509429MesuresTraficIlliciteEn.pdf/MesuresTraficIlliciteEn.pdf>.

⁵⁶ *Charter*, *supra* note 3.

Development of Policies and Requirements for Ingesting and Preserving Digital Records Into a Preservation System

Where to start?

Alicia Barnard

Abstract

In the light of the new Records/Archives Law in Mexico, it is mandatory to issue requirements and processes for agencies recordkeeping systems as well as for the development a digital repository for long-term preservation. For that there are already models, standards and basic requirements, widely accepted that may be reviewed and adapted for our environment. Moreover, good practices for digital records preservation require giving special attention to digital records appraisal activities. Thus, in order make recommendations for digital records appraisal policies for federal agencies, selected documents from national archives as well as research projects are being reviewed.

Author

Alicia Barnard is currently an independent records manager consultant. From 1990 to 2008, she was director of the Documentation Centre in the Mexican Federal Ministry of Health. In 1992, she received an “Archival Merit Diploma” and in 2000 “Acknowledgment to the Records and archival Tasks Diploma” from the General Archive of the Nation. From September 2005 to March 2006 she was a member of the CLAID Team funded by UNESCO to receive training on electronic records preservation within the InterPARES 2 Project. She was director of TEAM Mexico of the InterPARES 3 Project until spring 2008 and co-researcher until May 2012. She has published and presented widely on digital records management and preservation.

1. Introduction

Before the recent Federal Law of Records/Archives in Mexico last issued last January,¹ digital records were still not an issue to be taken into account. Although there were some unclear general guidelines for the Federal Government, record managers and archivists were and are indeed worried about this kind of records. One may say that until now information technology areas (IT) are ruling how the information is managed within Federal Agencies without any records/archival practices, although all information systems create, maintain and preserve information that comply or function as records since they are created as an instrument or by-product of an activity or as a reference, are unique, and have archival bond that relates with others of the same aggregation. Right now when we are about to have a new Administration nobody knows for sure how digital records from the current Administration are being delivered to the next one, or how are they kept or disposed of, or if there is enough information to assure that the systems where the information resides are trustworthy. Although computer safety regulations help the maintenance of information systems, we are aware that they complement records/archival practices, but it is not enough since nobody knows how data maintained in information systems is being disposed, or if said data is kept without any reason to do so, or if it is already compromised because of migration or other changes (technological, administrative or juridical) that might impact on its authenticity and accessibility.

¹ <http://www.diputados.gob.mx/LeyesBiblio/pdf/LFA.pdf> (Accessed August 28, 2012).

Well of the evils, the lesser. The above mentioned Federal of Records/Archives Law sets for the first time regulations for digital records, it stipulates that processes and technical tools are equivalent for both paper and electronic records, among them there are: capture (creation/integration), monitoring (tracking), use, creation of functional classification schemas, file integration, description (section, series, files), transference (disposal rules), maintenance, information access and personal data controls, appraisal and audit. Besides the Law disposes that guidelines for recordkeeping systems must be issued which should consider: a) To keep and preserve metadata created for the development of such recordkeeping systems; b) To include rules and measures to guarantee authenticity, security, integrity and availability of electronic records as well as those devoted to their management and control; c) to develop procedures to document updating, backing, migration, or other processes that affect the authenticity of electronic records as well as those juridical-administrative and technological changes in systems, software or devices and hardware that may also impact on the electronic records content.

Fortunately the Law dispositions have not forgotten digital records preservation, either for the creator's usability or for the historical and social values, the Law disposes that institutions must have a preservation system according to the specifications to be issued.

On these order of ideas, in addition to basic functionalities for a recordkeeping systems for agencies, in relation with records, National General Archives (NGA) must develop a digital repository together with policies, rules, guidelines or procedures related to appraisal and transference of records with values to be considered as heritage of our Nation.

As for a recordkeeping system there are already models, functionalities, requirements that are being analysed in order to adapt which is best for agencies. Those rules as well as the guidelines are being elaborated. Firstly, for next January the rules derived from the Law will be published as well as the basic policies and requirements for a recordkeeping system. In this task it was decided to select minimum requisites and review the further on them according to the experiences obtained in their instrumentation.

With respect to the transference/ingesting and preserving records to a digital repository, the OAIS model² already offers the main elements for transference/ingesting digital data. The model, provides with elements to consider for this process such as: agreements between the creator and the preserver in relation with the type of information to be transferred, formats, means of transmission and periodicity as well as the requirements to be established by the repository regarding the type of information that will keep and maintain, the kind of metadata needed, intellectual rights to be acquired, and so on. There are also criteria for digital records repositories certification.³ It is not the same with the appraisal process, there is not a single model or methodology to do so, but is also a fact that appraisal⁴ cannot be anymore carried without

² Consultative Committee for Space Data Systems (2012): Reference Model for an Open Archival Information System (OAIS), <http://public.ccsds.org/publications/archive/650x0m2.pdf> (Accessed August 28, 2012).

³ For digital repositories criteria check: The Center for Research Libraries and Online Computer Library Center, Inc. (2007): Trustworthy Repositories Audit & Certification: Criteria and Checklist. (2007). Version 1.0, http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (Accessed August 28, 2012); And also: International Standards Organization: ISO 16363:2012: Space data and information transfer systems -- Audit and certification of trustworthy digital repositories, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510 (Accessed August 28, 2012).

⁴ The definition of appraisal for this paper is the one of the InterPARES Glossary: "The process of assessing the value of records for the purpose of determining the length and conditions of their preservation," http://www.interpares.org/ip2/display_file.cfm?doc=ip2_glossary.pdf&CFID=2004855&CFTOKEN=47951380 (Accessed August 28, 2012).

specific policies methodologies or guidelines they are essential if we really want to maintain for future generations our heritage materials available.

Currently, for paper records Federal agencies, most of the states and local agencies already use functional classification schemas linked to disposition schedules, although it is perceived that these instruments are not always good or applied correctly, so first they must be reviewed for its application within a digital environment. As for an appraisal methodology GAN, as the agency responsible of authorizing records destruction or transference issued a reference guide for identifying paper records series with secondary value,⁵ it recommends agencies to carry out record series appraisal during the first and second phase of the life cycle, first for a quality study of the records series and a second appraisal for comparing other similar series created by other institutions, besides the guide offers certain general topics as well of a checklist that will lead to an approximate of certain documents with secondary values. GAN has also issued basic requirements for the transference of paper records to be preserved for the long-term. Notwithstanding there is still lack of a policies, guidelines or practices for appraising digital records.

The purpose of this study is to offer an approach of certain factors and recommendations to be taken into account for the development of appraisal policies in Mexico. It is based on the review of the main digital records appraisal policies or rules issued of the National Archives of Canada, United States, United Kingdom and New Zealand published in their official websites. Besides the InterPARES project materials related to appraisal will be taken into account, as well as the results of the recent conclusions obtained by the Digital Records Appraisal Group of the Latin American (including Spain) Records Appraisal Project (FIED).⁶ First, a synthesis of the documents reviewed is presented followed by an approach of what appraisal policies or guidelines in Mexico should consider.

2.1 National Library and Archives of Canada (LAC)

Document: Appraisal Methodology: Macro-Appraisal and Functional Analysis. Part B: Guidelines for Performing an Archival Appraisal of Government Records. 2000 reviewed on 2005⁷

Purpose:

The guidelines proposed are based on the Macro-Appraisal approach by means of functional analysis to get information about the environment where information is recorded such as its nature, structure, creation process and especially interrelationships with other information/records creators and users. The guidelines also considers at the end micro-appraisal to identify factors such as completeness and comprehensiveness, authenticity, uniqueness, relationship to other records, date and time span, extent,

⁵ Archivo General de la Nación (2009): Guía para la identificación de series documentales con valor secundario, <http://www.agn.gob.mx/menuprincipal/archivistica/pdf/GuiaIdentificacion21052012.pdf> (Accessed August 28, 2012).

⁶ The Digital Records Appraisal Group of the Latin American and Spanish Records Appraisal Project was partially funded from 2009 to 2012 by the International Council of Archives. The professors and practitioners that participated were from Argentina, Brazil, Colombia, Spain, Costa Rica, Mexico, Peru and Uruguay. There were different groups for specific topics related to appraisal, such as terminology, educational programs, sampling, methodologies and specific issues on digital records. Particularly, this last group was coordinated by Lluís-Esteban Casellas (in 2011) and Alicia Barnard (in 2010-2012) and integrated by Ma. Teresa Bermudez, Aida Cristina Oliveira, Andrés Pak Linares and Aída Luz Mendoza Navarro, <http://blogs.ffyh.unc.edu.ar/evaluaciondedocumentos/category/mariela-alejandra-contreras-argentina/> (Accessed August 28, 2012).

⁷ <http://www.collectionscanada.gc.ca/government/disposition/007007-1041-e.html>.

usability, physical condition. The guidelines were designed for both paper and digital records. Particularly, in relation with electronic data of program delivery or analysis systems, the guidelines recommend archivists to carry out:

Investigation of micro-data as well as longitudinal data-files built from said micro-data in order to (recognize or) confirm the role of regional and field data systems and their relationship with data sharing with superior systems.

2.2 National Archives and Records Management Administration (NARA) of the United States

Document: Appraisal Policy. September 2007⁸

Purpose:

The policy sets out the strategic framework, objectives, and guidelines that the National Archives and Records Administration (NARA) use to determine whether Federal records both, traditional and digital, have archival value. It establishes which are permanent records categories as well as specific guidelines for appraising certain categories of records.

The policy states that the authority for retention and disposition of the Federal Records is the National Archivist, although it also establishes that the process is carried out with interested parties and considers the point of view of the records creators. Besides the policy includes appendixes for special considerations such as the usability of electronic records that might require specific measures due to their technological capabilities in contrast with other records that are easy to maintain. There are also special considerations for certain types of records as observational data, environmental health and safety records or research and development records; said considerations are mainly directed to digital records.

2.3 The National Archives (TNA), United Kingdom.

Document: The National Archives Appraisal Policy (last updated August, 2004)⁹

Purposes:

- To develop a system of appraisal applicable to new environments created by digital records
- To ensure the continued transfer of paper records to TNA for at least the next 20 years.
- To ensure that appraisal for archival purposes selects records of highest archival value avoiding duplication.
- To provide appraisal methods for both digital and paper records as well as records created in any other medium.
- The policy states that an effective appraisal, mainly in the digital environment depends on good systems of records creation and business scheduling records.

⁸ <http://www.archives.gov/records-mgmt/initiatives/appraisal.html> (Accessed August 28, 2012).

⁹ http://www.nationalarchives.gov.uk/documents/information-management/appraisal_policy.pdf (Accessed August 28, 2012).

The policy was developed to deal with digital records, it admits that the traditional Grigg System issued in 1958 proved its effectiveness for paper records although specific changes should be made for appraising digital records. It states that an effective appraisal, mainly in the digital environment, depends on good systems of records creation and business scheduling records. Moreover, the policy considers macro-appraisal appropriate for digital records as an initial guide to identify public records of value for business and archival purposes, in order to understand the functions that create them. Also, it considers macro-appraisal as an aid to identify those records with potential archival value in file plans, as well as datasets and case files that overlap between departments. It also includes the need for developing generic archival appraisal guide for categories of records such as those produced by similar types of departments.

In relation with appraisal TNA has developed a various documents that help agencies with the process such as the Appraisal Report Template, General Guidelines for the Selection of Records, How to compile an appraisal report, Series level appraisal questionnaire.¹⁰

2.4 New Zealand Archives

Document: Appraisal Policy, September 2008¹¹

Purpose:

To support the Chief Archivist's decision making around appraisal of government records for the purposes of disposal.

The Policy briefly explains why appraisal is needed in context of government and local agencies and the purposes for determining which records are public. Also includes principles for good records practices to support appraisal decision making such as lawfulness, accountability and transparency, consistency and resources that should be taken into consideration together with the point of view of creators. Also, the policy establishes certain objectives that help to identify the archival significance of records and the responsibilities of the process for producers and preservers. The policy applies for any public record (electronic, paper or other) as stated in the Public Records Act.¹²

3. Main conclusions encountered in the documents reviewed

- Most Archival Institutions do have an appraisal policy.
- The policy main purpose is to support the National Archivist for disposition decision making actions.
- Policies also guide creators to identify records that should be preserved because of their values.
- Policies are focused on both traditional (paper, film, etc) and digital records.
- There are checklists to help the creator to identify those records with permanent values, as well as special considerations for scientific data, environmental data, etc.

¹⁰ <http://www.nationalarchives.gov.uk/information-management/guidance/a.htm> (Accessed August 28, 2012).

¹¹ http://archives.govt.nz/sites/default/files/appraisal_policy_0.pdf (Accessed August 28, 2012).

¹² <http://www.legislation.govt.nz/act/public/2005/0040/latest/DLM345537.html> (Accessed August 28, 2012).

- There are recommendations for macro-appraisal or functional appraisal although down to top appraisal is not left out.
- The main reason of an appraisal policy was for better appraisal practices regarding digital records and considers that an effective appraisal is supported by good systems of records creation and scheduling records.
- Recognition of the role of regional and field data systems and their relationship with data sharing with superior systems .
- The issuing guides for categories of records such as those produced by similar types of departments.

By reviewing the InterPARES Project proposals for appraisal there are four main issues to take into consideration for digital records appraisal.¹³

1. To carry out appraisal at the beginning of the life cycle or when designing the recordkeeping systems.
2. Assess and document authenticity. Digital records intangibility as well as technological obsolescence or its transmission in time and space are undoubtedly factors that compromise their authenticity. Therefore, assessment to identify elements related to its integrity (persons, dates, archival bond) and integrity (completeness and uniqueness) of records or the recordkeeping system where they reside must be carried out. Although one must accept that digital records of the creator are authentic unless the contrary is proved, we must trust on our digital heritage to be authentic and reliable for its preservation.
3. Determining feasibility of preservation. By identifying technical requirements for preservation, formats, as well as digital components such as metadata, content and context information will help the institution that is going to preserve the records for the long-term to understand costs of acquisition, taking into account that most of them are constant. There may be records that are developed in complex systems that might require either to postpone its transference in order to look for more resources or alliance with other institutions.
4. Monitoring appraised records. Those records already appraised for is long-term preservation require to be monitored and reappraised before they are transferred to an archive for preservation, in order to assess if the initial appraisal decision has not changed or if there is no damage in records and its components when there are changes or redesign of the recordkeeping system, or in functions within the organization, or in the status of the records information.

The InterPARES 3 Project with support of the International Council of Archives recently issued the series *Digital Records Pathways: Topics in Digital Preservation*.¹⁴ The series includes a module devoted to

¹³ Hackett, Yvette, Domain 3 Task Force, "Appendix 21: Preserver Guidelines – Preserving Digital Records: Guidelines for Organizations," [electronic version] in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records*, ed. Luciana Duranti and Randy Preston (Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008).
http://www.interpares.org/display_file.cfm?doc=ip2_book_appendix_21.pdf (Accessed August 28, 2012).

appraisal which reasserts that “principles for appraisal and preservation decisions should be embedded in all record-creating and recordkeeping activities.” The module proposes the following steps for conducting appraisal which is based on the above mentioned issues.

- **Compiling information.** This activity includes gathering contextual information of the record’s corpus to be appraised (juridical-administrative, provenancial, procedural and technological).
- **Assessing value.** The continuing value either for the creator due to legal, evidentiary or business reasons, or for cultural, historical and research purposes could be assessed either by top-down in terms of the records contexts or a bottom up approach by assessing their values.
- **Assess and document authenticity.** The InterPARES *Requirements for Assessing and Maintaining the Authenticity*¹⁵ was designed for this purpose in order to either establish if records may be ingested for preservation or if it is needed to carry out risk assessment when authenticity is compromised.
- **Determining the feasibility of preservation.** This activity implies to identify records elements to be preserved according to the system design and configuration, records elements that may be manifested in digital components in various ways (content, metadata and creation contexts). This will help to know about the system, the essential records manifestations, its metadata and context creation that will help to reconcile preservation requirements with preservation capabilities by assessing the institutions current and future preservation capabilities such as professional knowledge, expertise as well as IT infrastructure and financial resources.
- **Monitoring appraised digital records.** Once the appraisal decision has been taken, records that are going to be preserved for the long-term require to be monitored while they reside in the recordkeeping system of the creator in order to verify if the initial appraisal decision is still valid and terms and conditions of transfer or require a new appraisal process.

The Digital Records Appraisal group of the Latin-American (including Spain) Evaluation Project (FIED) focused its research on appraisal documents issued by different national archives as well as on international research projects or institutions, some of which were also reviewed for this study case.¹⁶ The main conclusions were:

- Appraisal as an intellectual activity is the same for any media.
- Appraisal must be carried out at the beginning of the lifecycle when records are controlled by a recordkeeping system

¹⁴ The InterPARES 3 Project, *Education Modules - Digital Records Pathways: Topics in Digital Preservation*, http://www.interpares.org/ip3/display_file.cfm?doc=Education-Modules_Digital-Records-Pathways.zip (Accessed August 28, 2012).

¹⁵ InterPARES: Authenticity Task Force (2002), “Appendix 2: Requirements for Assessing and Maintaining the Authenticity of Electronic Records” in *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, ed. Luciana Duranti (San Miniato, Italy: Archilab, 2005), 204–219. Online reprint available at http://www.interpares.org/book/interpares_book_k_app02.pdf.

¹⁶ The digital records appraisal group besides reviewed information from the website of the National Archives of Australia, the InterPARES Project also reviewed official websites of the Digital Curation Centre, Digital Preservation Europe, Inter-University Consortium for Political and Social Research, the UK Data Archive as well as the National Archives of Australia.

- Digital records appraisal cannot be dependent on technological factors, considerations on processes and contexts that impact the creation of digital records must be taken into account.
- Professional education, knowledge of the records environment and research are basic elements for good appraisal policy.

I would also like to mention the InterPARES Project Director Dr. Luciana Duranti's conclusions for digital records appraisal:¹⁷

- *Appraisal, different from selection, is entirely conditioned by the specific context.*
- *It requires a clear relationship between the creator and the designated preserver.*
- *Selection and appraisal must be based on trust.*
- *Appraisal must be clearly motivated on the basis of arguments that are not only being archival/diplomatic and technological, but also legal, ethical, and financial.*
- *The archivist must be all to all records, including the potential records that s/he should contribute to design.*
- *Appraisal must be carried out as soon as possible after creation.*
- *Appraisal must consider functions, records, systems.*
- *Appraisal must be carried out by a team of records professionals, legal, and financial experts, and IT professionals.*
- *Appraisal would serve the creator, researchers and society at large.*
- *Appraisal should be carried out in the environment in which they reside at any given time.*
- *Appraisal should be carried out every time the system's creator is upgraded or changed. Once the records are acquired the selection process and the appraisal is definitive.*

Thus, the Digital Appraisal Group of FIED developed the following recommendations when designing appraisal policies:

- To develop and implement classification schemas, records schedule plans, migration, refreshment and updating procedures before transferring records to a digital repository for its preservation. This in order to assure integrity, reliability and usability of digital records.
- To take into consideration problems associated with records hosted outside the servers and to establish the development of recommendations in this respect.
- To elaborate guidelines for metadata retention regarding the appraisal decision and its link to the record during the retention period, as well as for protection against unauthorized disposition.
- To establish monitoring strategies to identify juridical-administrative, provenancial, procedural and technological changes of those records already appraised for its preservation.
- To consider that information creator systems must comply with recordkeeping procedures and processes in order to identify and establish retention periods, to provide periodical information related with the systems updates or changes, transfer or destruction activities according to schedule plans, and documentation related to the system functionalities

¹⁷ Duranti, Luciana, keynote address, "The Appraisal of Digital Records: The Return of Diplomats as a Forensic Discipline," International Conference on "The E-Discovery Challenge: Digital Wasteland or Digital Oasis. An Interdisciplinary Approach to Managing Records – Archives, Law and Evidence," 3-4 November 2010. Singapore, Singapore, http://www.interpares.org/display_file.cfm?doc=ip3_canada_dissemination_cs_duranti_singapore_2010a.pdf.

- To consider the elaboration of guidelines for the control and elimination of records produced or included in websites, and when necessary, to adapt schedules plan in order to include series created for the website when these are not already considered.
- To consider appraisal of information as by product of scientific data analysis models.

Besides, the Digital Appraisal Group established certain requisites and issues to consider when appraising digital records.

The information gathered lead to conclude that there is enough information for designing a general appraisal policy with special attention to digital records which deserve utmost attention due to the constant risk they face because of the technological environment that compromises their authenticity, thus severely questioning if they are trustworthy for the their preservation. Thus the following topics are of relevance to be considered.

- The role and responsibilities of creator and preserver for the appraisal process.
- The statement of the main governmental, cultural, historical and societal reasons for continued preservation.
- Check lists that may help the creator in the decision making process.
- The appraisal process must be carried out at the beginning of the life cycle or when designing the recordkeeping system for both digital and paper records.
- The development of specific policies, guidelines or procedures:
 - For carrying out the appraisal process. These will make distinction of traditional records (paper or other) and digital ones.
 - For specific disposition criteria in relation with generic series that are created and maintained in different agencies mainly for those created and maintained in digital environments.

Particularly, the appraisal policy must consider the following topics for digital records.

- Functional appraisal methodology as an initial approach although down-top appraisal should not be ignored.
- The main activities to carry out digital appraisal, such as:
 - Assessment and documentation of authenticity.
 - Identification of digital components of records (content, metadata and contextual information).
 - Determining feasibility of preservation.
 - Monitoring records already appraised.
 - Developing transfer-ingest plans.
- Development of risk analysis strategies or diplomatic analysis when it is presumed that records have high secondary values but authenticity is compromised because of the lack of structural or formal components or because they are not captured in a recordkeeping system.
- Appraisal carried out by multidisciplinary groups, based on local and national regulations.

Notwithstanding, the availability of policies, methodologies, procedures or requirements to be adapted or adopted for a good appraisal, as the main step to continued preservation of high value records, our digital heritage in Mexico will remain in danger if authorities of the highest level forget that archival/diplomatic practices are urgently needed to create, maintain and preserve digital records. Also it is important for the authorities to recognize that the challenge for preserving digital records is not anymore an isolated task carried out by records managers or archivists they should be accompanied by other professionals since the best solutions are multidisciplinary shared by creators, records managers and preservers and other stockholders. In these tasks IT also plays an important role, but not the only one as it is still perceived by authorities in Mexico.

Where Light in Darkness Lies

*Preservation, Access and Sensemaking Strategies for the Modern Digital Archive*¹

Jason R. Baron¹ and Simon J. Attfield²

¹US National Archives and Records Administration; ²Interaction Design Centre, Middlesex University, USA

Abstract

The second decade of the 21st century finds institutions around the world increasingly having to cope with the matter of how to preserve electronic records in response to litigation, regulatory and compliance demands. In the public sector, existing approaches to email preservation in particular (based on past litigation) run the gamut from continued reliance on print to paper strategies; to deployment of disaster recovery backup tapes as default recordkeeping systems; to forms of electronic recordkeeping that continue to rely on end users performing records management functions; and most recently, to automated email archiving. This paper argues that with greater encouragement and adoption of automated capture methods, archivists and historians should understand both the limitations of present-day search techniques, as well as the need to extract “meaning” out of what are increasingly vast amounts of electronic records, especially through consideration of new methods drawn from the disciplines of sensemaking and visual analytics.

Authors

Jason R. Baron serves as Director of Litigation for the US National Archives and Records Administration, and is an Adjunct Faculty Member at the University of Maryland, College of Information Studies. Mr. Baron formerly held the positions of trial attorney and senior counsel in the U.S. Department of Justice. He has also served as a Visiting Scholar at the University of British Columbia, participated in InterPARES, and co-founded the US NIST TREC Legal Track. He is a recipient of the Emmett Leahy Award, recognizing his career contributions in the field of records and information management. Mr. Baron received his degrees from Wesleyan University and the Boston University School of Law.

Simon Attfield is a Senior Lecturer in the Interaction Design Centre, Middlesex University, UK. His research lies in the area of understanding how people work with information, processes involved in individual and collaborative sensemaking and implications for interactive systems design. With Ann Blandford he is co-author of the book *Interacting with Information*, part of the Morgan Claypool series of Synthesis Lectures on Human-Centered Informatics. He received a B.A. in Philosophy and a BSc. in Experimental Psychology from Sussex University, and a Ph.D. in Human Computer Interaction from University College London.

1. Introduction

During the past 20 years, with the growth of computer networks and interconnectivity in the workplace,² email has become a worldwide phenomenon, transforming the lives of individual employees. The

¹ This article represents a reworking of themes and materials presented at the 8th European Digital Preservation Conference (ECA 2010) in Geneva, at I-CHORA 5 (2010) in London, as well as the DELOS conference (2007) in Rome. The authors wish to thank Richard Cox, Richard Pearce-Moses, Mary Jo Pugh, and Gary M. Stern for their comments on prior drafts, as well as Luciana Duranti for her overall support. The views expressed here are solely the authors, however, and do not purport to represent the official position of any institution with which they are affiliated.

replacement of secretaries with personal computers has, for better or worse, turned each office worker into a *de facto* record keeper in public bureaucracies, and to a lesser extent, in private firms and corporations as well. Now, at the start of the second decade of the 21st century, public and private institutions around the world are increasingly coping with the consequence of having empowered all staff with email as a communications tool, viewing email as a corporate necessity. Paradoxically, *de facto* - modern email archives remain a major source of risk, given the minefield represented by modern day litigation, investigations, compliance and regulatory measures, all in search of evidence in the form of a candid “smoking gun.” And yet, the email universe continues to expand exponentially, giving rise not only to short term information governance challenges, but also profound issues regarding access to the contents of those communications—including the desire to find relevant messages as well as in making sense of the collection as a whole.

Public sector institutions have deployed variety of policy and technology solutions as “preservation” policies: continued reliance on print to paper strategies; deployment of disaster recovery backup tapes as default recordkeeping systems; electronic recordkeeping that relies on end-users to “tag, drag and drop” individual email communications into an archive; and most recently, tentative adoption of automated capture methods known as “email archiving.” All these technological solutions, other than “print to paper,” more or less rely on automated searching to provide access to email archives in their electronic form. By now, however, well-established limitations on the efficacy of keyword searching leave open for resolution how best alternative strategies may be employed for extracting “meaning” from vast amounts of email, i.e., how to perform information retrieval to extract the “needles” of meaning from the ever growing e-haystack.³

There is, however, a fundamental “records management” reality that needs to be confronted and overcome. Even two decades into the deployment of email, and even in the highly networked world we find ourselves in, the reality of the “end user’s” experience in managing and preserving electronic records has not fundamentally changed. In particular, in the United States at present, only a few public sector institutions have instituted successful, comprehensive, enterprise-wide electronic content management, even in the face of sustained criticism that the state of records management in the federal sector in the U.S. approaches “chaos.”⁴ Change is now in the air, however, as most notably evidenced in new records management mandates from President Obama and U.S. Archivist David Ferriero.

² Martin Hilbert and Priscilla Lopez, “The World’s Technological Capacity to Store, Communicate, and Compute Information,” *Scienceexpress* (Feb. 10, 2011) (in 2007, humankind was able to store 290 exabytes and the rate is increasing by 23% a year), accessed October 9, 2012, <http://www.sciencemag.org/content/332/6025/60>.

³ See: “The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods Used in E-Discovery,” *Sedona Conference Journal* 8 (Fall 2007), www.thesedonaconference.org/publications.

⁴ Citizens for Responsibility and Ethics in Washington (CREW), “Record Chaos: The Deplorable State of Electronic Record Keeping in the Federal Government,” (2008), <http://www.scribd.com/doc/49055979/Record-Chaos-Report>; U.S. Government Accountability Office, “Federal Records: National Archives and Selected Agencies Need To Strengthen E-Mail Management,” GAO-08-742 (2008), <http://www.gao.gov/new.items/d08742.pdf>.

2. A New Day for Managing Public Sector Email & A Challenge

On November 28, 2011, President Obama issued a memorandum on Managing Government Records for the heads of all Executive branch agencies in the U.S., stating that he was beginning a government-wide “effort to reform records management practices.”⁵ He recognized that:

Decades of technological advances have transformed agency operations, creating challenges and opportunities for agency records management. Greater reliance on electronic communication and systems has radically increased the volume and diversity of information that agencies must manage. With proper planning, technology can make these records less burdensome to manage and easier to use and share. But if records management policies and practices are not updated for a digital age, the surge in information could overwhelm agency systems, leading to higher costs and lost records.

In stating that “proper records management is the backbone of open government,”⁶ the President directed that the Archivist of the United States, working with other senior officials, develop a Records Management Directive (“Directive”) that addresses modern day records challenges, including with respect to email, social media, and cloud based data sets.

On August 24, 2012, the Archivist of the United States issued the planned for Directive, a document that represents an inflection point in the history of archival institutions worldwide. The Directive boldly sets out that by 2019 all permanent electronic records of the U.S. Government are to be managed and preserved in electronic form, for eventual transfer and accessioning at the National Archives and Records Administration (“NARA”).⁷ As a corollary, the Archivist also directed that *all* email records (both permanent and temporary in nature), be managed in an accessible electronic format by the end of 2016. Specifically with respect to email, the Directive goes on to say:

Email records must be retained in an appropriate electronic system that supports records management and litigation requirements (which may include preservation-in-place models), including the capability to identify, retrieve, and retain records for as long as they are needed.⁸

The Directive separately calls for NARA and others to work with private industry “to produce economically viable automated records management solutions,” including “advanced search techniques,”⁹ and to “embed records management requirements into cloud architectures.”¹⁰ These promising mandates at long last give a much needed impetus to finding satisfactory electronic archiving solutions for email and other forms of electronic records. They represent a profound commitment to true digital government: that what is born digital should be preserved digitally, at least insofar as the data, information and records represent what is deemed appropriate to permanently preserve the memory of the 21st century.

But these aspects of the Archivist’s directives represent a challenge to archivists, as well as to all those seeking access to archives (including lawyers, historians, and researchers)—given the new reality that we are living in a world of vast public sector digital archives that are expected to be sustained on an

⁵ See: <http://www.whitehouse.gov/the-press-office/2011/11/28/presidential-memorandum-managing-government-records>.

⁶ Ibid.

⁷ Id., Part I, § 1.1.

⁸ Id., § 1.2.

⁹ Id., Part II, § A3.

¹⁰ Id., § A4.

indefinite basis into the long-term future. Compliance with the new Directive requires the management and ultimately long-term preservation of permanent record emails in digital form. Given this mandate, archivists in public sector institutions such as NARA also have a responsibility to be good stewards of these growing digital archives in providing access. Thus, beyond resolving difficult preservation issues, greater attention also needs be paid to the conduct of search and sensemaking within these large digital collections, for purposes of best ensuring or optimizing future access as expeditiously as possible, within the constraints of the law, for the benefit of future generations.

In the United States, due to litigation involving White House email that began on the last day of the Administration in Ronald Reagan (in January 1989), presidential records in the form of email are deemed permanent public records and have been preserved.¹¹ By 2017 there may be as many as one billion White House emails cumulatively in the legal custody of the Archives.¹² The White House email archiving experience against the backdrop of continuing litigation lasting over two decades amounts to a tale both fascinating and cautionary, with a mixed bag of lessons.¹³ That this archive has been continuously maintained over two decades, across a variety of proprietary platforms—with whatever their admitted defects—serves to indicate that automated solutions to email archiving are indeed feasible.

Yet, paradoxically, with limited exceptions all email that has come into the National Archives to date remains behind closed doors, essentially locked away in the digital dark. Litigation and Freedom of Information Act requests provide limited forays into these dark places, but they do not represent the kind of systematic review that is needed to open these archival collections in a logical way. NARA's presidential libraries are overwhelmed with lengthening access queues due to both litigation and filings under the Freedom of Information Act. However, because the collections do contain sensitive and privileged information—everything from social security numbers to medical conditions to prior criminal records and the like—they cannot easily be made open to the public without further archival processing. This in turn means that large segments of these email repositories exist as *de facto* digital dark archives, awaiting future systematic opening in whole or in part determined by the longest period covered by existing federal privacy laws and practices. Under such circumstances, the default condition on these archives is that they will not become open to the public until 75 or more years after the end of the current Administration.¹⁴

¹¹ See: *Armstrong v. Executive Office of the President*, 1 F.3d 1274 (D.C. Cir. 1993).

¹² See: George L. Paul and Jason R. Baron, "Information Inflation: Can The Legal System Adapt?" *Richmond Journal of Law and Technology* 13, no. 3 (2007): 1-41, <http://jolt.richmond.edu/v13i3/article10.pdf> (estimating 1 billion White House emails in the legal custody of NARA by 2017).

¹³ See: Jason R. Baron, "The PROFS Decade: NARA, Email and the Courts," in *Thirty Years of Electronic Records*, ed. Bruce I. Ambacher (Lanham, MD and Oxford: The Scarecrow Press 2003), chap. 6; David Bearman, "The Implications of *Armstrong v. Executive Office of the President* for the Archival Management of Electronic Records," *American Archivist* 56, no. 4 (1993): 674-89; GAO Report 01-446, "Clinton Administration's Management of Executive Office of the President's E-mail System (April 2001), <http://www.gao.gov/products/GAO-01-446>; Citizens for Responsibility and Ethics in Washington, "The Untold Story of the Bush White House E-mails," (August 2010), <http://citizensforethics.org/bush-white-house-ignored-warnings-about-email>; Myron Groover, "The White House E-Mail Destruction Scandal of 2007: A Case Study for Digital Heritage" (paper presented at the UNESCO Conference: The Memory of the World in the Digital Age, Vancouver, B.C., 2012).

¹⁴ See: NARA regulations at 36 C.F.R. 1256.56(a)(2) (2012) (NARA policy potentially restricts information in personnel, medical or other records revealing information of a highly personal nature regarding living individuals that inter alia "relates to events less than 75 years old.").

In light of all of the above, a commitment to sustainable digital archives necessitates a parallel commitment to new ways of thinking: first, about the burden on end-users in performing records management; second, about the need for automating the segregation or classification of permanent records apart from the disposable or transitory, and third, about the opportunity to use advanced search techniques to extract meaning and make sense of the coming vast digital collections.

3. Past Failed Approaches to Email Management

We first need to declare an end to the era of end-users assuming the burden of records management. Since the late 1980s, with the universal rollout of PCs in offices and at personal workstations, all end-users have become *de facto* records managers, which may have been theoretically achievable in a world of word processing, spreadsheets, and the *occasional* email of substantive importance. However, the reality of the present day workplace is vastly different: each end user is drowning in information, only a portion of which in the public sector constitute long-term temporary or permanent records.

Except in certain exotic and rarified environments, print-to-paper appears to be an utterly failed recordkeeping paradigm. Substantial rates of individual end-user noncompliance with retention policies calling for the printing out of email are the norm in any organization that will admit to such in a survey (or under oath). The U.S. Government Accountability Office reported in 2008 on a small survey of 15 senior officials at various federal agencies, eight of whom “did not consistently conform to key requirements in NARA’s regulations for email records, such as filing them in appropriate recordkeeping systems.”¹⁵ The simple truth: with rare exception, for the past twenty years, few professionals (and even fewer lawyers) consistently or comprehensively have printed out all emails of record for placement in traditional office filing systems.

But other more digital-centric recordkeeping approaches that nevertheless rely on end-users have similarly met with less than success. Strategies that rely on ad hoc actions on the part of individual end-users to archive email into personal folders do not constitute comprehensive recordkeeping of value to the organization as a whole, and these methods flunk the current test under federal e-recordkeeping guidelines for what constitutes an electronic recordkeeping system.¹⁶ At the same time, some percentage of public sector organizations continue to store email on disaster recovery backup tapes with minimal or no recycling involved, as their electronic recordkeeping “solution.” Agency officials are often oblivious to the distinction between “archiving” and “recordkeeping” functions, as well as to the difference between each of those and saving data to disaster recovery backup tapes—thus policies relying on backup tapes may not be seen by them as inherently problematic. Email and other e-records “stored” on these forms of backups, however, have been physically streamed in no logical order and must be “restored” by a labor-intensive process of remounting sets of backups from a single day or session, to extract desired files from reconstructed user accounts. Disaster recovery backup tapes are not databases or logical repositories of

¹⁵ See: GAO Report 08-742: 4. See also: NARA, “Records Management Self-Assessment: An Assessment of Records Management Programs in the Federal Government,” 2009, <http://www.archives.gov/records-mgmt/resources/self-assessment.html>; SRA International, Inc., “Report on current Recordkeeping Practices with the Federal Government,” (Dec. 10, 2001), <http://www.archives.gov/records-mgmt/pdf/report-on-recordkeeping-practices.pdf>.

¹⁶ See: 36 C.F.R. 1236.20 (requiring records disposition functionality and shared access).

information and cannot be easily searched without great time, money and effort.¹⁷ That is why NARA states in its federal recordkeeping regulations that backups “must not be used as the agency electronic recordkeeping system.”¹⁸

A cottage industry of records management applications (RMAs) by way of software products and services also exists. These products and services provide organizations with the capability of empowering end-users with greater control of the ability to perform electronic recordkeeping. The most visible standard for e-recordkeeping is the 5015.2 Standard (now in version 3 form) employed by the U.S. Department of Defense since 1997, which sets out several dozen functional requirements for RMA software to meet in order to demonstrate compliance with federal recordkeeping standards.¹⁹ A significant number of software products have been certified as 5015.2 compliant, and NARA has encouraged their use.²⁰

For all of the support given during the past decade of this “first wave” of 5015.2 compliant RMA applications, they have not been widely deployed to date within government agencies. Part of the problem may be residual difficulty in scaling up the software that comes out of the box, to meet the needs of particular large-scale institutions and enterprises. However, what we perceive to be the much larger issue is the transactional demands that this form of software makes on end-users, the vast majority of whom do not consider records management a top priority as they go about their daily workplace tasks. RMAs modeled on 5015.2 version 3 normally require users to select how individual emails and other e-records are to be managed from drop-down screens. Often times, institutions have not thought through how they would simplify their existing stock of legacy records schedules, so as not to impose a bewildering array of choices on users indicating the series in which individual electronic objects are to be placed. The result in some cases has been failed pilot projects, where users simply rebelled, or in the more usual case, underutilized the RMA application due to the transactional burden involved. The jury measuring the success of RMA applications is still out, however: it may well be that the product lines need further maturation, that the end-user experience will be easier, and that this form of e-recordkeeping solution might eventually become more pervasive than it is currently.

Anecdotal experience with DoD Standard 5015 RMA applications has shown that many individuals do not reliably save, tag, drag or drop most of their email into electronic folders set up with well-intentioned archival purposes in mind. The compliance rate starts off low in this regard, and is invariably getting lower—simply due to massive volumes of information now confronting us all. Having to perform extra keystrokes, no matter how few, on any substantial percentage of electronic communications during the work day, means the equivalent of having to pay a transactional toll per communication, in terms of lost time, energy, and productivity. Few individuals wish to pay the price on a consistent and comprehensive basis, hence, a world of incomplete if not haphazard recordkeeping.

¹⁷ See generally: Grant J. Esposito and Thomas M. Mueller, “Backup Tapes, You Can’t Live With Them and You Can’t Toss Them: Strategies for Dealing with the Litigation Burdens Associated with Backup Tapes Under the Amended Federal Rules of Civil Procedure,” *Richmond Journal of Law and Technology* 13, no. 3 (2006): 1-26, <http://law.richmond.edu/jolt/v13i3/article13.pdf>.

¹⁸ 36 C.F.R. 1236.20(c) (2012).

¹⁹ See: <http://www.archives.gov/records-mgmt/policy/joint-interopability-letter.html>; see also: <http://jitec.fhu.disa.mil/recmgt/>.

²⁰ NARA, “Endorsement of DoD Electronic Records Management Application (RMA) Criteria Standard,” *NARA Bulletin* 2003-03 and 2008-07, versions 2 and 3, respectively, <http://www.archives.gov/records-mgmt/bulletins/2003/2003-03.html> and <http://www.archives.gov/records-mgmt/bulletins/2008/2008-07.html>.

Aside from the inherent problems of backup tapes, the Achilles heel for the remaining above-described methods should be plain: that is, the extent to which these methods rely on individuals to perform recordkeeping (“self-archiving”) at the desktop. Given the reality of the sheer volume of email and other electronic records coming across the transom every day, coupled with the transactional cost of compliance, sole dependence on software that requires diligence by end-users is tantamount to institutions whistling past the graveyard, at least with respect to their immediate e-discovery obligations, and with downstream recordkeeping implications absent strong training and enforcement programs in place. In our view, agencies have a responsibility to look for alternative strategies in light of these real-world considerations.

4. The Promise of Email Archiving Through Automated Capture

As distinct from the above approaches, it is now increasingly apparent that software technology has advanced significantly to the point that we can speak in terms of something truly new: automated email archiving for the enterprise. Such a notion of comprehensive capture of email by automated means conjures up Borgesian visions²¹ of near-infinite library repositories—a dream to some, a nightmare to others, including perhaps appraisal archivists grounded in longstanding notions of what constitutes the “judicious” disposition of records.²² At least today, email archiving software starts with the promise of the comprehensive capture of email, not just as selected by users, but as pulled from the underlying proprietary email system on a continual basis. The concept of “automated email archiving” is predicated on the notion that 100% of email traffic is captured or routed into some form of online or near-line electronic environment. Without the type of fanfare accompanying iPhones and other 21st Century new age gadgetry, this new form of software and services nevertheless has gained a wide audience, and is currently being piloted in a variety of both public and private sector institutions as a way to manage risk, including the risk associated with having to comply with litigation demands for “all” e-communications.

Automated email archiving is wonderfully simple in its conception: email, and whatever else, is captured from whatever email proprietary system or email store is in use on the online server, with the electronic object then placed in a separate database either near-line or offline to be preserved in its native proprietary form under whatever rule set is proscribed. The electronic objects are therefore accessible from a central location. Indeed, the system may well be set up that the user is able to “see” or otherwise freely access such archived email in the separate database as if the email still resided in his or her original in-box (through the use of what is referred to as an email “stub”). From all of the above discussion, it should be clear that the value of such systems for e-discovery purposes is apparent, in lowering the risk of lost records while increasing accessibility through means of expedited searches.²³

As NARA’s bulletin on the subject of email archiving pithily states, “Email archiving applications typically require little to no action on the part of the user to store or manage the email records.”²⁴ NARA recognized that benefits include (i) more efficient storage of email because it is moved from a distributed

²¹ Jorge Luis Borges, “The Library of Babel,” in *Labyrinths: Selected Stories and Other Writings* (Penguin, 1971), 51.

²² Title 44 of the U.S. Code, Section 2902(5) (listing “judicious preservation” as one of the objectives of records management).

²³ See: Wikipedia, “Email archiving,” accessed October 9, 2012, http://en.wikipedia.org/wiki/Email_archiving.

²⁴ NARA, “Guidance concerning the use of e-mail archiving applications to store e-mail,” *NARA Bulletin* 2011-03 (December 2010), <http://www.archives.gov/records-mgmt/bulletins/2011/2011-03.html>.

network of servers and desktop applications into one place; (ii) enhanced search capabilities for content germane to subpoenas, FOIA requests, and e-discovery requests; and (ii) assistance in backup and disaster recovery.²⁵ NARA's bulletin goes on to note, however, that "[e]-mail archiving is a relatively new technology and is still developing," and that agencies must appropriately configure and implement records management controls when using the application.²⁶

We believe, however, that solutions already exist in the marketplace that meet some level of the recordkeeping controls that NARA would wish to see in place. For example, consistent with existing records retentions schedules, an agency may decide to implement an automatic rule "tagging" all email created or received by designated senior officials as "permanent." Remaining email from all other agency staff could be categorized as presumptively "temporary" in one or more "big buckets" or folders within an email archiving scheme corresponding to existing record series. Alternatively, a similar tagging rule could be employed for all email created or received by designated components that routinely generate high-level policy or other sensitive documents worthy of long-term preservation.

An agency might elect any number of "add-on" measures, to bring greater nuance to the archiving scheme. These might include simple steps aimed at allowing for users to create individual folders within the email archive, for the limited purpose of ensuring that some measure of "virtual foldering" takes place akin to traditional records management in the paper world. For senior staff with designated permanent records, agencies concerned about over-inclusion of email of a personal or truly ephemeral nature could allow for staff to delete emails from the in-box for a limited period of time (e.g., 60 or 120 days), with any emails remaining then automatically captured as permanent under the "auto-archiving" rule in effect. Conversely, mid- or lower level staff who do create "permanent" records could be allowed a manual override of the "temporary" designation, so as to "opt-in" for records preservation on individual or designated categories of email sent or received.

In marketing email archiving solutions, much has been made of the notion that the software also generally allows for greater records management functionality. One of the multiple reasons why organizations implement email archiving is said to be to meet litigation, regulatory, and/or business records retention requirements, by enabling easier searches of stored email. On closer examination, there appear to be no perfect, scalable solutions yet to truly automating a large organization's email by content, even using techniques borrowed from the world of information retrieval, data mining, and artificial intelligence. Some vendors in the marketplace are, however, touting the ability to have systems intelligently "auto-classify" or "auto-categorize" documents based on some measure of machine learning from examples provided by users themselves.²⁷ Such techniques would serve as proxies to more traditional means of recordkeeping that those of us born in the 20th Century are used to in the workplace. Even now, however, the technologies available in the marketplace go a serious way towards meeting the long-anticipated goal of some archivists in being able to adopt more sophisticated forms of macro-appraisal and "institutional functional analysis," for reliably segregating content by function or activity of an organization,²⁸ and thus are worthy of serious further examination.

²⁵ Ibid., p. 2.

²⁶ Ibid.

²⁷ See Jason R. Baron, "Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and The Current State of E-Discovery Search," *Richmond Journal of Law and Technology* 17, no. 3 (2011): 1-33, <http://jolt.richmond.edu/v17i3/article9.pdf>.

²⁸ Terry Cook, "Archival Appraisal and Collection: Issues, Challenges, New Approaches," Special Lecture Series, University of Maryland and to NARA Staff, 21-22 April 1999, <http://www.mybestdocs.com/cook-t-nara-990421->

5. Limitations of Present-Day Search Methods and A Path Forward

To date, the majority of products and services touting electronic archiving solutions appear to have put their eggs in the back-end search basket, rather than implementing more traditional front-end records management solutions. This suggests that search capabilities are robust enough to replace all other notions of provenance, original order, and the need for any form of granular auto-categorization by record series type, features of a repository that archivists have traditionally expected and demanded. In light of the claims being made, it remains important to accurately measure success in performing robust information retrieval. This is, however, a rapidly evolving field, and alternative solutions to Boolean searching being deployed in present day litigation may yet have important implications, including for members of the archival profession.

To the extent email archiving schemes rely on sorting by means of present-day “search” technologies to perform records management-like classification, they are of course on somewhat shaky ground. As is well known in the information retrieval community,²⁹ the idea that any form of text retrieval algorithm can reliably parse text such as email, based on content alone, into that which is valuable and worth keeping, while setting aside that which is ephemeral and to be disposed, still is viewed as a hard problem—especially if the goal is perfect sorting of records into 20th Century-style, highly granular records schedule buckets. There are, however, a number of products and services in the e-discovery market that take advantage of clustering algorithms and “predictive” analytics, in ways that would greatly advance the auto-categorization task.

This subject has been explored at length in connection with e-discovery issues, through publications including The Sedona Conference® *Best Practices Commentary on the Use of Search and Information Retrieval Methods Used in E-Discovery*.³⁰ The Commentary describes at length the limitations involved in present-day search methods based on keywords and Boolean operators, and suggests that lawyers (and others) should be looking in the short term to alternative forms of search methods. The latter methods now include forms of supervised learning (widely known today in the legal community as “predictive coding”), that makes use of the coding of seed sets of documents, coupled with use of statistical or probabilistic forms of searching based on mathematical clustering models, for the purpose of enabling software to in turn code the vast majority of documents in a given repository.³¹ Given the exponential increases in volume of electronically stored information, lawyers are increasingly taking advantage of the well-known gains in efficiency in relying on these software-assisted methods, which in turn utilize “iterative” feedback loops involving greater transparency in divulging statistical samples to those inquiring about relevant documents.³²

2.htm; Terry Cook, “Electronic Records, Paper Minds: The revolution in information management and archives in the post-custodial and post-modernist era,” *Archives and Manuscripts* 22, no. 2 (1994): 300-328; see also: Christopher Tomer and Richard J. Cox, “Electronic Mail: Implications and Challenges for Records Managers and Archivists,” *The Records & Retrieval Report* 8, no. 9 (1992): 7 (referring to a “macro-appraisal/documentary probe” strategy for email).

²⁹ For citations to the relevant literature, see: Douglas W. Oard and William Webber, “Information Retrieval for E-Discovery,” *Foundations and Trends in Information Retrieval* 6, no. 1 (forthcoming in 2012): 1-144, accessed October 9, 2012, <http://ediscovery.umiaccs.umd.edu/pub.html>.

³⁰ See n.3, *supra*.

³¹ Baron, “Law in the Age of Exabytes.”

³² Paul and Baron, “Information Inflation,” 50.

For example, in the prominent *da Silva Moore* case decided in 2012, a U.S. magistrate judge held that the state-of-the-art in advanced search techniques had progressed to the point where the Court could “bless” the use of the form of supervised learning known as “predictive coding.”³³ The court determined the use of predictive coding was appropriate considering “(1) the parties’ agreement, (2) the vast amount of electronically stored information to be reviewed (over three million documents), (3) the superiority of computer-assisted review to the available alternatives (*i.e.*, linear manual review or keyword searches), (4) the need for cost effectiveness and proportionality . . . ; (5) the transparent process proposed by [defendants].” The opinion included at its end an extraordinarily detailed “joint protocol” worked out by the parties, setting out the construction of seed sets, sampling, and the seven rounds of iterative training of the system for purposes of classifying documents as responsive or nonresponsive. The opinion points to a future of judicially-blessed litigation engagements where supervised learning in lieu of keyword searching are routinely utilized.

Additionally, over the past half decade the limitations of competing information retrieval methods in a legal context have been explored as part of what is known as the “TREC Legal Track,” an international research project sponsored by the U.S. National Institute of Standards and Technology, which ran between 2006 and 2011. The goal of the legal track was to evaluate how alternative forms of search methods perform in a “real-life” situation involving real document databases (consisting to date of seven million tobacco litigation documents, plus the Enron email collection), and a set of hypothetical topics which could have been the subject of e-discovery demands. In other words, the track coordinators wished to draw comparisons as between how well Boolean methods perform against fully automated alternatives proposed and used by scientist-participants in the track. The results of the second year of the Legal Track were eye-opening (at least to lawyers, as opposed to information retrieval researchers): only 22% of the total number of relevant documents were found by traditional keyword or Boolean search methods; 78% were found by all other search methods combined.³⁴ Results from subsequent years have validated the above numbers while also exploring the efficacy of greater use of human-in-the-loop solutions, as well as supervised learning methods.³⁵

The research and analysis represented by the TREC Legal Track points to the difficulty in making very strong claims regarding the efficaciousness of how reliably searches, conducted against a large electronic archive with many millions or billions of objects embedded within it, can successfully classify or segregate information categories. As stated, these methods are rapidly improving, however. Arguably, parsing documents into categories of “relevance” is more difficult than attempting to match documents into relatively broad recordkeeping categories (by series, for example), although any attempt to categorize documents based on content analytics alone will achieve less than perfection. (Of course, some measure of success can be obtained if an organization is willing to set big-bucket rules to designate as permanently valuable “all” records generated by particular individuals or components within the enterprise, without regard to content, as discussed, *supra*.) The bottom line is that while we should remain cautious about naively buying into overly robust assertions regarding the efficacy of searches, there is rapidly emerging

³³ *Moore et al. v. Publicis Groupe SA*, 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012) (Peck., M.J.), *aff’d*, 2012 WL 1446534 (S.D.N.Y. April 26, 2012) (Carter, J.)

³⁴ See: Jason R. Baron, “Discovery Overload,” *Law Technology News* (2008): 36, reprinted at <http://commonscolld.typepad.com/eddupdate/2008/01/edd-showcase-di.html>; Stephen Tomlinson, et al., *Overview of the TREC 2007 Legal Track*, http://trec.nist.gov/pubs/trec16/t16_proceedings.html.

³⁵ Oard and Webber, “Information Retrieval for E-Discovery”; See: TREC Legal Track Overview papers, <http://trec-legal.umiaccs.umd.edu/>.

progress in the area of search methods and technologies that archivists and lawyers would be well served to remain aware concerning. The promise of the future is that by employing greater use of sampling and iterative techniques, coupled with new forms of supervised learning, e-record archives can potentially be filtered for other purposes—including for performing macro appraisal and disposition, and for determining where resources are to be allocated to find subsets of electronic repositories that may be made available for access in the nearer term.

There are many approaches to search and information retrieval, only some of which have been explored by those in the legal domain. We turn next to the issue of extracting meaning or “sensemaking” from email archives, an area which holds great promise as it becomes better understood.

6. Sensemaking and the Email Archive

In recent years interest in the study of sensemaking has increased, and one area where this has been evident is in the study of people working with electronic information. If we can understand how people make sense of things and how that can best be supported, we are in a better position to design systems to help sensemaking along, and so help people to engage effectively with increasingly large amounts of electronic content. As both a research area and a practical issue, this interest can be said to be motivated by a need for perspectives that link people’s moment-by-moment interactions with information, such as searching, gathering, extracting, and structuring, with the internal processes of theorizing, interpreting and understanding.

A number of theoretical perspectives on what sensemaking is and what happens during sensemaking have emerged.³⁶ A theme that commonly runs through these accounts is the observation that sensemaking involves a reciprocal interplay between bottom-up exposure to information on the one hand, and the top-down interpretation, structuring or theorizing about information on the other. Each occurs during sensemaking, and each affects the other. Interpretation gives meaning to information and even defines what counts as information (as opposed to ‘noise’). But it is the information which gives support to or challenges the interpretation, perhaps forcing it to be modified or abandoned in favor of another. Sensemaking has been described as a process of placing stimuli into some kind of framework (e.g., a mental narrative, a model of others’ motives and intentions, a spatial “map,” etc.), which then allows us to “comprehend, understand, explain, attribute, extrapolate and predict.”³⁷

The co-dependence between information and interpretation in sensemaking presents something of a paradox. Appreciating what information is relevant to an endeavor depends on interpretation which itself depends on information. We equate this aspect of sensemaking to the idea of the hermeneutic circle.

³⁶ See, e.g., Karl Weick, *Sensemaking in Organisations* (London; Sage, 1995); Brenda Dervin, “An Overview of Sense-making Research: Concepts, Methods, and Results to Date” (paper presented at the International Communications Association Annual Meeting, Dallas, May, 1983); Gary Klein, et al., “A Data-frame Theory of Sensemaking” in *Expertise Out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making, Pensacola Beach, Florida, May 15-17, 2003*, ed. Robert Hoffman (Lawrence Erlbaum Associates Inc., 2007), 113-155; Peter Pirolli and Stuart Card, “The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis” (paper presented at the International Conference on Intelligence Analysis, McLean, VA, May 2-3, 2005), <https://analysis.mitre.org/proceedings>.

³⁷ William H. Starbuck and Frances J. Milliken, “Executives’ Perceptual Filters: What They Notice and How They Make Sense,” in *The Executive Effect: Concepts and Methods for Studying Top Managers*, ed. Donald C. Hambrick (Greenwich, CT, JAI Press, 1988), 35-65.

Hermeneutics concerns the theory (or art) of interpretation.³⁸ Originally concerned with the interpretation of written texts, contemporary hermeneutics has come to include the analysis of interpretive processes in general. According to the hermeneutic circle, parts of a message can only be understood in terms of an understanding of the whole, and yet an understanding of the whole can only arise from an understanding of the parts.³⁹ Take a sentence—individual words, which individually are susceptible to multiple interpretations, have their interpretation fixed through an interpretation of the sentence as a whole. And yet, the attributed meaning of the sentence whole depends upon how each individual word is interpreted. This scales-up—the interpretation of a message, such as an email, depends upon some theory of the context in which the message originated (e.g., time, culture, language, personalities, roles, motivations, recent activities and relationships)—and yet an understanding of these depends upon the interpretation of many other such messages.

We illustrate this with an example from a study into the ways in which lawyers makes sense of large collections of emails during some e-discovery investigations. Attfield and Blandford⁴⁰ reported an interview study conducted with corporate lawyers who were engaged in a e-discovery investigations. One group of lawyers was investigating potential fraud in the contracts that a company had with another. They conducted keyword searches over a set of recovered documents (primarily emails) and the results were retrieved so that they could manually review them. However, an interpretation of actions in the emails depended upon an understanding of the broader temporal backdrop against which it occurred. For example, the judgment of whether a communication between two parties could be interpreted as bid rigging would depend on when the communication had taken place in relation to a bid lifecycle. But this context was initially unknown, and needed to be constructed from many other similar emails. People may meet, exchange information or even money, but the meaning of these actions, and, important for this example, their meaning in legal terms is fixed by a broader context. Hence, a bootstrapping process must occur at great time and cost. As one senior lawyer said, “the scope of what you’re trying to do is immense and you’re having to define it as you go along.”

Sensemaking involves moving from the part to the whole and back, gradually building an understanding. For the philosopher Schleiermacher, the impasse of the hermeneutic loop can be broken by an initial cursory reading of an entire text followed by detailed readings of specific parts and relating these to the whole. Schleiermacher, however, did not have large email collections to investigate. A cursory reading of thousands to millions and even tens of millions of documents is not possible.

There is, however, an emerging technology which offers an opportunity for addressing the hermeneutic loop. Visual analytics is the science of analytical reasoning supported by interactive visualizations. At its heart is the idea of interactive graphical representations of large datasets which are, in principle, easy to interpret and provide context from which to engage in more detailed investigation. Visual analytics has at its core interactive visualizations which utilize the high bandwidth of the human visual system to enable us to draw insights from large, complex datasets. Computerized visualization is not new, of course, but with the concept of Visual analytics it becomes part of a more holistic study of computer-supported sensemaking. The form of the solution is to perform automated analysis (of some

³⁸ “Hermeneutics” and “interpretation” have a shared etymology. See: Lawrence K. Schmidt *Understanding Hermeneutics* (Stocksfield, UK: Acumen, 2006).

³⁹ Schmidt, *Understanding Hermeneutics*.

⁴⁰ Simon Attfield and Ann Blandford, “Discovery-led Refinement in E-discovery Investigations: Sensemaking, Cognitive Ergonomics and System Design,” *Artificial Intelligence and Law* 18, no. 4 (2010): 387-412.

type) over a document collection and to use the results to structure visual displays that offer insights into the underlying collection. Visual overviews abstract away from data, prioritizing some aspects of underlying content and hiding others.

Visual analytic tools are a relative newcomer to the e-discovery stage, but given the importance of email to this domain, the interactive visual exploration of email has become a significant area of interest.⁴¹ A number of approaches to the visualization of email collections have been offered. It is not our aim to review them here and the interested reader might refer to a review by Lemieux and Baron.⁴² We do, however, refer to one example by way of illustration of the approach.

A research system called Enronic by Heer⁴³ is shown in figure 1. Enronic is an example of a network graph visualization tool. The system takes an email collection and graphically displays connections based on who sent emails to whom. An advantage of generating social networks from email archives lies in the possibility of calculating and social network metrics and using these as display variables. For example, Enronic calculates *connectivity* (often referred to as *degree centrality*) for each

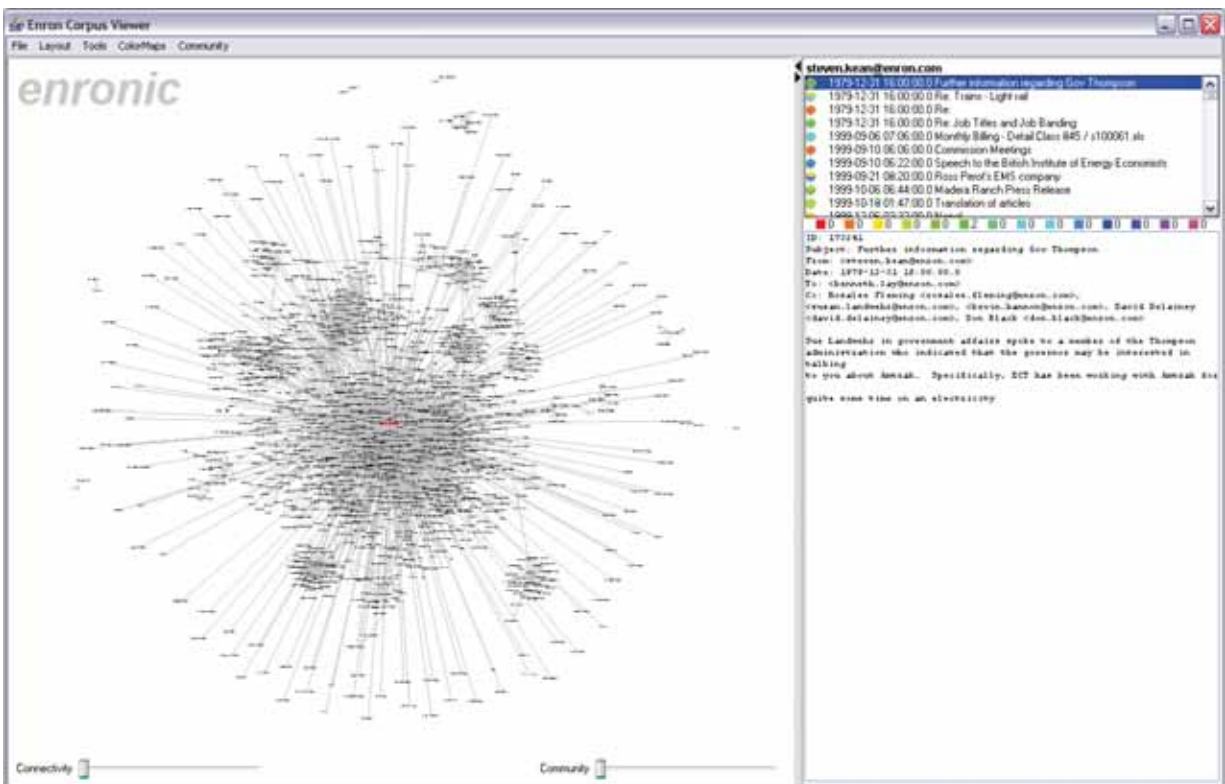


Figure 1. Enronic displays connections inferred from a set of emails, in this case the Enron archive. It provides an interactive visual representation of the network and *connectivity* (*degree centrality*) potentially indicating a person's power or influence within the network.

⁴¹ Victoria L. Lemieux and Jason R. Baron, "Overcoming the Digital Tsunami in E-Discovery: Is Visual Analysis the Answer?" *Canadian Journal of Law and Technology* 9, no. 1 (2011): 33-48.

⁴² Ibid.

⁴³ "Exploring Enron: Visualizing AMLP Results," accessed October 9, 2012, <http://hci.stanford.edu/jheer/projects/enron/v1/>.

node. This is a measure of the number of connections emanating from a node and can potentially indicate a person's power or influence within the network. Various forms of filtering are typical within Visual analytic systems. Enronic allows the user to filter the display for nodes with higher connectivity. Users can also apply an algorithm to hierarchically cluster the network into community substructures and interactively view them in terms of different levels of hierarchical agglomeration.

Displaying the social network implicit within an email collection is one kind of overview representation; it involves exploiting email metadata. As Lemieux and Baron⁴⁴ observe, other common forms include timelines (also based on metadata) and representations of content such as galaxy views in which similar documents are clustered together. From the perspective of a given investigation, one type of view may not be enough. And so we anticipate that systems that support people in making sense of email archives (for litigation purposes or otherwise) will incorporate multiple interactive visualizations of different types. Further, multiple, tightly coupled visualizations may be used concurrently, with users cross-referencing between, or sequentially, with different visualizations suited to different tasks in an analytic chain and at different scales of magnitude.

Sensemaking is frequently slow and painstaking. However, Visual analytic tools in conjunction hold out some promise for constructing scalable methods for making sense of large collections of documents such as emails. Just as these tools are becoming more significant for lawyers, so too we expect that they will become more significant for archivists, historians and others interested in uncovering hidden secrets that email archives may have to offer.

7. The Archivist of the Future and The Modern Digital Archive

Archivists and records managers need to confront reality in understanding that institutions are facing extraordinary pressures to adopt such automated email archiving schemes. With the encouragement of the Archivist's Directive, and in the absence of prior institutional motivation to move forward with other forms of automated or electronic recordkeeping (e.g., 5015.2 compliant solutions), a present day vacuum exists: those institutions (including most of the U.S. government) that still live in a paper-based world with respect to their default, official recordkeeping system will likely leapfrog to automated capture schemes involving email archiving before they make large expenditures in deploying other forms of enterprise-wide software. The inevitable result: automated email archiving scheme will, more likely than not, also be perceived to be an institution's *state of the art records management scheme*, and it will be exceedingly difficult to convince end-users to continue any semblance of other means of recordkeeping, in the face of the knowledge that email archiving exists.

This brings us to the final, most profound challenge that current day automated email archiving schemes pose to archivists and historians: the desire on the part of institutions to save everything for the here and now, but to delete (almost) everything after a prescribed period of years (e.g., all non-permanent records). Automated email archiving has a kind of "July 4 fireworks" aspect. In the United States, July 4th—Independence Day—is a day traditionally celebrated with fireworks, which are of course both highly illuminating and highly ephemeral in nature. The profound problem posed by automated email archiving is that it has the potential to turn out to contain no "archives" of any sort—rather, it may have been conceived to date by some individuals and institutions as simply a short-term fix consisting of

⁴⁴ Ibid.

information that may be wholly deleted after a set term of years (without the need for any form of segregation of “permanent” record holdings). Such an archives, despite the vast amount of information contained within, would ironically leave no permanent mark. Like fireworks, petabytes or exabytes of electronic records would alight the sky for a brief period of years, and then, in terms of the institutional memory of the public sector in the 21st century, all would grow dark. The Archivist’s Directive challenges the federal sector in the United States to address the need to extract “meaning” from federal databases by the end of this decade, by segregating permanent records for preservation in electronic form. However, all institutions confront similar issues in separating the permanent from the ephemeral.

The gauntlet being thrown down to the archival profession should be clear: records managers and archivists of all stripes demand that deployment of any such capturing schemes be accompanied by some recognition on the part of the institution that, if they are to morph into the *de facto* recordkeeping scheme for the institution, the business rules of the institution should well call for categories of “permanent” and “temporary” records being culled out and properly segregated. To the extent the IT community is empowered to build such structures, others within the organization should fairly see them for what they are, and demand front-end thinking in the procurement cycle as to the long-term records management implications of deployment. In doing so, archivists and records managers may yet be able to import “the best of” 5015.2 electronic recordkeeping into a model for automated email archiving, with greater use of auto-categorization methods, to come up with workable hybrid model.

This paper argues that we should acknowledge the obvious: end-users cannot cope with recordkeeping demands imposed by email, and thus, the end of the end user as records manager is near. But what does the implication of automated electronic archiving mean for future records managers and archivists? What role do they have to play when retention environments are “fully automated”? Records managers and archivists hold the potential to exercise enormous influence on how electronic records are managed and preserved, including ensuring that electronic archiving systems with recordkeeping functionality are sufficiently in place to allow for contemporaneous capture of permanent records in electronic forms, for immediate copying and eventual migration to the archival institution itself. This is an idea whose time has come.

In light of the reality of litigation, we believe that the “least worst” solution is for institutions to put into place automated capture solutions for email for purposes of information governance and overall risk management. We also see a need for future archivists to accept and prepare for the technological future that confronts them. This is a future where intelligent filters and automated rules-based systems replace records series and records categorization as practiced in the 20th Century; and where both front-end and back-end solutions based on notions of artificial intelligence, data mining, content and visual analytics, and the like are increasingly employed to separate wheat from chaff, the permanently valuable from the flotsam and jetsam of more ephemeral forms of records. Through such means archivists will truly make sense of the modern email archive.

Writing in *The American Archivist*, Richard Pierce-Moses has spoken to the demands on the archivist of the future in a similar vein:

[A]rchivists should become as comfortable working with digital records as they are working with traditional media. Instead of pen and paper, we will work with cursor and keyboard. Instead of sorters, we will work with sorting algorithms. Rather than weeding, we will filter. With few exceptions, all archivists will need to know what we now call technical skills as the vast majority of contemporary and future records are and will be digital. Different archivists will need different technical skills No doubt some

archivists will continue to specialize, but their specializations will be specific to the digital arena: databases, image and audio formats and metadata, but also user interfaces, search systems, and digital preservation.⁴⁵

The volume, growth, and complexity of electronic records is confronting the legal and business sectors with serious challenges, and as we have argued elsewhere, a critical juncture has been reached where 20th century ways of doing business, at least in terms of responding to discovery demands in lawsuits, no longer suffice.⁴⁶ Litigation and other forms of external access demands may be drivers of change, but they need not drive organizations off the cliff in terms of adopting reasonable solutions for the management of records both in line with short-term business needs and for history's larger purposes. Archivists and records managers, working with lawyers, IT staff, and business executives, have the opportunity to collaborate in coming up with appropriate business models that satisfactorily take into account all of these competing demands and priorities. This is not an easy road for the archivist of the future, but it is a necessary one.

Finally, we have no doubt that changing technology in this area will render some of what is said here rapidly obsolete. We trust, however, that there is more than a little value in attempting to stake out a position on these subjects, even if in so doing one must recognize that we are all engulfed inside a vortex of transformative change. The change is in our workplaces as well as in our collective cultures, with respect to what constitutes the adequate capture in modern digital archives of permanently valuable electronic records, both for business purposes as well as for the sake of history's greater illumination.⁴⁷

⁴⁵ Richard Pierce-Moses, "Janus in Cyberspace: Archives on the Threshold of the Digital Era," *The American Archivist* 70, no. 1 (2007):18; See also: Philip C. Bantin, "Strategies for Managing Electronic Records: A New Archival Paradigm? An Affirmation of Our Archival Traditions?" (n.d.), <http://www.indiana.edu/~libarch/ER/macpaper12.pdf>.

⁴⁶ Paul and Baron, "Information Inflation."

⁴⁷ "So ere you find where light in darkness lies, your light grows dark by losing of your eyes." William Shakespeare, *Love's Labour's Lost*.

Strengthening the Regulatory Framework in a Digital Environment

A Review of Archives Legislation

Elaine Goh

School of Library, Archival and Information Studies, The University of British Columbia, Canada

Abstract

This paper outlines some of the challenges posed to archival legislation in Commonwealth countries regarding the creation, management, and preservation of records in increasingly complex digital environments. The author argues that archival science can contribute to the strengthening of current archival legislation, so as to address issues in the digital environment. The author also proposes future areas of empirical research on archival legislation.

Author

Elaine Goh is a doctoral student at the University of British Columbia. Her research interest is on archival legislation in Commonwealth countries and on organizational culture and behaviour. Elaine is a graduate research assistant of the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) Project. Prior to starting her doctoral program, Elaine worked as the Assistant Director of Records Management in the National Archives of Singapore.

1. Law and Technological Development

Legislation is one of the “medium(s) through which law is expressed.”¹ One approach in terms of understanding how law is expressed is the instrumental and functionalist perspective. This perspective argues that the rules of law should be imbued with certain qualities or attributes. For example, the rules of law should be stable in terms of comprising “determinate requirements that people should consult before acting and not retrospectively establish legal obligations.”² The law should have the quality of clarity and precision, so that those involved in the implementation and enforcement of the law understand the rules and are able to apply the law consistently.³ The law should also be imbued with the quality of “foreseeability” in terms of its ability to predict future events, in order to guide human actions and behavior.⁴ Despite the fact that most legislations aim to have “precise prescriptions”, such prescriptions may not necessarily be a “faithful description of any state of affairs but a complex ideal that is even more complex to realize.”⁵

¹ Wim Voermans, “Quality of EU Legislation Under Scrutiny: What Kind of Problem, by what Kind of Standards,” in *Quality of Legislation - Principles and Instruments - Proceedings of the Ninth Congress of the International Association of Legislation (IAL) in Lisbon, June 24th-25th, 2010*, ed. Luzius Mader and Marta Tavares de Almeida (Germany: Nomos, 2011), 35.

² Naomi Choi, *Rule of Law* (Thousand Oaks, California: Sage Reference Online), accessed 22 July 2012, <http://knowledge.sagepub.com/view/governance/n475.xml?rskey=XDEOlq&row=4>.

³ Jorge Miranda, “Law, Rule of Law and Quality of Law,” in *Quality of Legislation: Principles and Instruments - Proceedings of the Ninth Congress of the International Association of Legislation in Lisbon, June 24th-25th, 2010*. (Germany: Nomos, 2011), 27.

⁴ Ibid.

⁵ Roderick A. Macdonald, “Legislation and Governance,” in *Rediscovering Fuller: Essays on Implicit Law and Institutional Design*, ed. Willem J. Witteveen and Wilbren Van der Burg (Amsterdam: Amsterdam University Press,

In the legal literature, there has been extensive coverage with regard to the inability of the law to keep up with technological change. According to Moses, “as technology gives rise to new possibilities, and people engage in new forms of conduct, the law continues to be directed to solving old problems and is unable to keep up with the modern world.”⁶ Consequently, “technological change can make the law become unclear and it can make law that was previously unobjectionable become subject to criticisms.”⁷ Another risk caused by the inability of the law to address technological developments is that “courts and legal practitioners are faced with applying decades—and even centuries—old definitions and principles to radical technologies not even conceived of less than twenty years ago.”⁸

One area where one can witness the inability of existing laws to address challenges in the digital environment are records related legislations, which is defined as legislation that “deals with records or information generally such as evidence legislation, which is not in connection with a specific legislated activity.”⁹ This can be illustrated by the recent review of the *Evidence Act (1997)* in Singapore. The Minister of Law in Singapore in the second reading of the Evidence (Amendment) Bill explained that when the act was first introduced in 1996, a “cautious approach” was taken in terms of admissibility of computer output as evidence.¹⁰ The Minister described the approach as a “cumbersome process not consonant with modern realities” and stressed that “computer output evidence should not be treated differently from other evidence.”¹¹ The Minister’s comments resonated with a consultation paper by the Singapore Law Academy.

The Academy claimed that that the *Evidence (Amendment) Act 1996* was developed on the premise that the certification of electronic records was based on a client-server network model, where responsibility of records is vested with the systems administrator. However, businesses are now adopting a delegated responsibility model, where responsibility for the computing system and business operations are now distributed among the client, information technology vendors and with other service providers.¹² The Academy observed that “in such computing and business models, it will be hard to identify the party or organisation responsible for the reliability of the electronic evidence.”¹³ The statement made by the Academy outlines some of the major challenges faced by archivists and records professionals in terms of ensuring the long-term trustworthiness of records over space and through time, particularly through the

1999), 288; Naomi Choi, *Rule of Law* (Thousand Oaks, California: Sage Reference Online, accessed 22 July 2012, <http://knowledge.sagepub.com/view/governance/n475.xml?rskey=XDEOlq&row=4>).

⁶ Lyria Bennett Moses, “Agents of Change: How the Law ‘Copes’ with Technological Change,” *Griffith Law Review* 20, no. 4 (2011): 763

⁷ Lyria Bennett Moses, “Adapting the Law to Technological Change: A Comparison of Common Law and Legislation,” *University of New South Wales Law Journal* 26, no. 2 (2003): 396.

⁸ Gregory E. Perry and Cherie Ballard, “A Chip by any Other Name would Still be a Potato: The Failure of Law and its Definitions to Keep Pace with Computer Technology,” *Texas Tech Law Review* 24 (1993): 799.

⁹ Jim Suderman, Fiorella Foscarini, and Erin Coulter, “International Research on Permanent Authentic Records in Electronic Systems (InterPARES 2 Project) Policy Cross-Domain: Archives Legislation Study Report,” 2 September 2005, 4, accessed 5 September 2012,

http://www.interpares.org/display_file.cfm?doc=ip2%28policy%29archives_legislation_study_report.pdf.

¹⁰ Singapore. Second Reading Bills. Evidence Amendment Bill, 14 February 2012, accessed 5 September 2012, http://sprs.parl.gov.sg/search/topic.jsp?currentTopicID=00076883-WA¤tPubID=00076904-WA&topicKey=00076904-WA.00076883-WA_3%23id-6e0461e8-8588-49d0-b05e-fc6c2d596955%23.

¹¹ Ibid.

¹² Daniel Seng and Sriram Chakravarthi, *Computer Output as Evidence* (Singapore: Singapore Academy of Law, 2003), 80, accessed 5 September 2012,

http://www.lawnet.com.sg/legal/ln2/comm/PDF/Computer_Output_as_Evidence.pdf.

¹³ Ibid.

use of cloud computing. Cloud computing poses several risks that can compromise security and adversely affect the governance framework for the management and preservation of records. These risks reflect the need of developing and strengthening existing archives acts in relation to records management functions. As argued by Chasse, “the silence of case law does not justify the silence of legislation. The impact of electronic technology upon law and practice, and its rapid change, should lead to the conclusion that legislation is needed before the law is demonstrably inadequate.”¹⁴

2. Objectives

This paper outlines the major challenges posed by digital technologies, particularly in a cloud computing environment, and discusses how such an environment can compromise the trustworthiness of records. The paper puts forth the position that a strong archival legislation can help to institute adequate controls for the proper creation, maintenance and preservation of records. In addition, the paper highlights some of the current weakness and inadequacies of certain archives acts, and argues that archival science can bridge the gap between the current archival legislation and its ability to address issues in the digital environment. Finally, the paper concludes by recommending some possible areas of research relating to archival legislation.

3. Challenges in the Cloud Computing Environment

Cloud computing is defined as a “model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers storage, applications and services) that can be rapidly provisioned and released with minimal management effort or cloud provider interaction.”¹⁵ Cloud computing is not a new technology per se, but it is regarded as a novel service delivery model, designed to bring about better economies of scale and scalability of information technology services and infrastructure.¹⁶ Tucker, in a CTO Roundtable discussion explained, “cloud computing is not so much a definition of a single term as a trend in service delivery taking place today.”¹⁷

The National Institute of Standards and Technology in the United States has outlined three service models.¹⁸ First, the Cloud Software-as-a-Service (SaaS) allows consumers to access the application system, and the cloud service provider will provide the necessary infrastructure and application capabilities. Second, the Cloud Platform-as-a-Service (PaaS) involves the cloud provider supplying the necessary infrastructure, operating environments, and tools for the development of specifically created or

¹⁴ Ken Chasse, “Electronic Records as Documentary Evidence,” *Canadian Journal of Law and Technology* 6 (2007): 142, accessed 5 September 2012, http://cjlt.dal.ca/vol6_no3/chasse.pdf.

¹⁵ Peter Mell and Timothy Grance, *The NIST Definition of Cloud Computing - Recommendations of the National Institute of Standards and Technology* (Gaithersburg: National Institute of Standards and Technology, [2011]), vi, accessed 5 September 2011, <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>.

¹⁶ Chris Rose, “A Break in the Cloud? The Reality of Cloud Computing,” *International Journal of Management and Information Systems* 15, no. 4 (2011): 59; Jared A. Harshbarger, “Cloud Computing Providers and Data Security Law: Building Trust with United States Companies,” *Journal of Technology Law and Policy* 16 (2011): 232.

¹⁷ Mache Creeger, “CTO Roundtable: Cloud Computing,” *Communications of the ACM* 52, no. 8 (August 2009): 52.

¹⁸ Peter Mell and Timothy Grance, *The NIST Definition of Cloud Computing - Recommendations of the National Institute of Standards and Technology* (Gaithersburg: National Institute of Standards and Technology, [2011]), 4, accessed 5 September 2011, <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>.

acquired applications by the consumer. Finally, under the Infrastructure-as-a-Service (IaaS), model, the consumer does not physically manage the infrastructure, such as computers and networks, but the consumer has the flexibility to control both the operating environment, such as the type of operating system and database, and type of application systems.¹⁹ Regardless of the model adopted by the user, cloud computing is essentially a distributed service delivery model, which involves the outsourcing of information and communications technology (ICT). This has implications in terms of the management and preservation of digital records.

One challenge in cloud computing is concerns raised with regard to the security of the recordkeeping infrastructure. Cloud providers may make changes to the computing infrastructure, the operating environment, and/or the application implementation, in order to deal with issues relating to usage load and storage. These technological changes may inadvertently affect the security of the recordkeeping system.²⁰ The nature and extend of this effect is partly dependent on the specific type of service model. For example, the use of SaaS as a cloud service model, i.e., web-based collaboration tools such as Google Apps, means that the responsibility for network, infrastructure security, and application code security primarily vests with the cloud provider.²¹ In an IaaS service model, the cloud provider has control over the security of the computing facility, whereas the cloud user has control and responsibility over the application code.²²

There are also transborder jurisdictional issues with regard to the control of records in a cloud computing environment, since records may be stored on data centres in different locations. Certain countries may not necessarily conform to the data protection and privacy related legislations and policies of the countries of the record creators.²³ There might also be an “unknown number of copies of the same digital document in different iterations across different jurisdictions” which “could affect the identification of relevant data for criminal proceedings.” Multiple copies of records also pose problems for records retention, since it is “common for service providers to replicate records for multiple backup, sending copies to sites in different locations or even different jurisdictions” and “this can mean that time-expired records are not properly deleted from every server held in every site.”²⁴ The accuracy, reliability and authenticity of records is at risk if the identity of the records are altered, should there be a lack of audit trails of the recordkeeping system and if there are no proper procedures to ensure proper delineation

¹⁹ Ibid., pp.2-3; Chris Rose, “A Break in the Cloud? the Reality of Cloud Computing,” *International Journal of Management and Information Systems* 15, no. 4 (2011): 61-64.

²⁰ Scott Paquette, Paul T. Jaegar, and Susan C. Wilson, “Identifying the Security Risks Associated with Governmental use of Cloud Computing,” *Government Information Quarterly* 27, no. 3 (2010): 245-253.

²¹ Peter Mell and Timothy Grance, *The NIST Definition of Cloud Computing - Recommendations of the National Institute of Standards and Technology* (Gaithersburg: National Institute of Standards and Technology,[2011]), 2-3, accessed 5 September 2011, <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>; Stephen Mason and Esther George, “Digital Evidence and ‘cloud’ Computing,” *Computer Law & Security Review* 27, no. 5 (2011): 525.

²² Peter Mell and Timothy Grance, *The NIST Definition of Cloud Computing - Recommendations of the National Institute of Standards and Technology* (Gaithersburg: National Institute of Standards and Technology,[2011]), 4, accessed 5 September 2011, <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>.

²³ Miranda Mowbray, “The Fog Over the Grimpen Mire: Cloud Computing and the Law,” *SCRIPTed* 6, no. 1 (2009): 135-136, <http://www.law.ed.ac.uk/ahrc/gikii/docs3/mowbray.pdf>; Stephen Mason and Esther George, “Digital Evidence and ‘cloud’ Computing,” *Computer Law & Security Review* 27, no. 5 (September 2011): 525-526.

²⁴ Australasian Digital Recordkeeping Initiative, *Advice on Managing the Recordkeeping Risks Associated with Cloud Computing*. 2010: Council of Australasian Archives and Records Authorities, 29 July 2010, 10.

of responsibility between the cloud provider and the user.²⁵ Issues with regard to the ownership, custody of records are also at risk, particularly if service providers merge and dissolve. In addition, there are challenges in exporting and migrating records across various platforms, particularly if record owners decide to change cloud providers.²⁶ These data migration challenges can be due to the incompatibility of data storage and transmission formats between vendors, or the sheer scale and complexity of the data to be migrated.

Given such challenges, it is clear that outsourcing of ICT services does not mean an outsourcing of risks. Record creators cannot entirely trust records in a cloud computing environment and it will require far more than authentication technologies to enable such records to be trustworthy. When archivists speak of trust in records, we refer to the “accuracy, reliability and authenticity of records.”²⁷ As “targets of our primary trust,” we can place our trusts on records, provided that there are adequate controls on the policies, procedures, mechanisms governing the record’s creation and maintenance stage.²⁸ However, trust is also a “three part relation that is grounded in the truster’s assessment of the intentions of the trusted with respect to some action.”²⁹ Hardy adds that “A trusts B to do X” and that “trust depends on the context.”³⁰ This relational aspect of trust relations between records creators and government agencies can be enforced through understanding these two parties as “agencies of accountability.”³¹ These agencies “provide insurance of trustworthy conduct, by putting pressure (facilitating, controlling or sanctioning) on persons, roles, institutions or systems that are the targets of our primary trust.”³² The trust relations between the record creator and the national archives can be enforced through articulating their roles and responsibilities as articulated in the archives law. Given that the trustworthiness of records are at risk in a digital environment and given the importance of recordkeeping in society, the law can help in sustaining the “trust relationship” between the creating agency and the preserver by “(taking) over those areas in which there is significant value at stake.”³³

4. Analysis of Archival Legislation in Commonwealth Countries

Archives researchers have critiqued the archives acts in their respective countries as being ineffectual and reactive with regard to management of digital records. Archives acts are described as being “weak”, “outdated, “old and inconsistent.”³⁴ Most archival legislation in Commonwealth countries based their

²⁵ Katharine Stuartand and David Bromage, “Current State of Play: Records Management and the Cloud,” *Records Management Journal* 20, no. 2 (2010): 220-221.

²⁶ Barclay T. Blair, “Governance for Protecting Information in the Cloud,” *Association of Records Managers and Administrators International*, 2010, HT 4, <http://www.arma.org/HotTopic/HotTopic910.pdf>.

²⁷ ICA International Terminology Database, accessed 20 July 2012, <http://www.web-denizen.com/>.

²⁸ Piotr Sztompka, *Trust: A Sociological Theory* (UK: Cambridge University Press, 1999), 47-48.

²⁹ Hardin, Russell, *Trust and Trustworthiness* (New York: Russell Sage Foundation, 2002), xx.

³⁰ *Ibid.*, p. 9.

³¹ Piotr Sztompka, *Trust: A Sociological Theory* (UK: Cambridge University Press, 1999), 47-48.

³² *Ibid.*

³³ Hardin, Russell, *Trust and Trustworthiness* (New York: Russell Sage Foundation, 2002), 64.

³⁴ Chris Hurley, “From Dust Bins to Disk-Drives and Now to Dispersal: The State Records Act 1988 (New South Wales),” *Archives and Manuscripts* 26, no. 2 (November 1998): 390-409; The National Archives. “Proposed National Records and Archives Legislation - Proposals to Change the Current Legislative Provision for Records Management and Archives - Consultation Paper,” (2003) <http://collections.europarchive.org/tna/20081023125241/http://www.nationalarchives.gov.uk/documents/policy-consultation.pdf> (accessed 20 July, 2012).

archival legislation on the UK Public Record Act of 1958, an act written for a paper based records environment and unable to meet the challenges posed by the digital environment.³⁵ In fact, the National Archives in UK consultation paper on a proposed archival legislation writes,

The public sector needs a legislative framework which will assure the accuracy and comprehensiveness of the records it makes. Keeping records that can serve as evidence of an organisation's policies, procedures, actions and decisions, and associated matters of governance, accountability and propriety cannot be left to chance.³⁶

One weakness of the current archival legislation in Commonwealth countries is that it does not stipulate the recordkeeping roles and responsibilities of government agencies as records creators. The act only states that government agencies should seek responsibility from the national authority before destroying public records. Most archival legislation in Commonwealth countries stipulate the role of the national archives in terms of acquiring, preserving and promoting access to archival records. In other words, the archival legislation fulfils a constitutional function in terms of establishing an institution.³⁷ The archival legislation in Canada and Singapore also states that the national archives should play an advisory role in conducting or facilitating the development of a records and/or information management programme in the government. In Singapore, the act even state that the national archives "shall advice public officers concerning standards and procedures pertaining to the management of public records."³⁸ However, archival legislation in Canada and Singapore is notably silent with regard to the shared lines of responsibilities and the accountability structures, with regards to recordkeeping and preservation, between record creators and the national archives. As such, there is no means of verifying that records creators and preservers fulfill their responsibilities according to accepted professional standards.

Establishing lines of responsibilities and an accountability framework through an archival legislation helps to preserve the trustworthiness of records. Through such lines of responsibilities and accountability frameworks, the creating agency is given the "primary responsibility for their reliability and authenticity while they are needed for business purposes," while the archives is accorded "responsibility for their authenticity over the long-term."³⁹ Moreover, as the reliability of records is dependent on the "completeness of the record's form and the amount of control exercised on the process of its creation," archival legislation should specify that government agencies must exercise due diligence

³⁵ Michael Roper and Laura Millar, ed., *A Model Records and Archives Law* (United Kingdom: International Council on Archives and International Records Management Trust, 1999), http://www.irmt.org/documents/educ_training/public_sector_rec/IRMT_archive_law.pdf.

³⁶ The National Archives. "Proposed National Records and Archives Legislation - Proposals to Change the Current Legislative Provision for Records Management and Archives - Consultation Paper," (2003), 7, accessed 20 July, 2012, <http://collections.europarchive.org/tna/20081023125241/http://www.nationalarchives.gov.uk/documents/policy-consultation.pdf>.

³⁷ Wim Voermans, "Quality of EU Legislation Under Scrutiny: What Kind of Problem, by what Kind of Standards," in *Quality of Legislation - Principles and Instruments - Proceedings of the Ninth Congress of the International Association of Legislation (IAL) in Lisbon, June 24th-25th, 2010*, ed. Luzius Mader and Marta Tavares de Almeida (Germany: Nomos, 2011), 35.

³⁸ *National Heritage Board Act*, Singapore Statutes Online, 1993, no. 13, <http://160.96.185.113/aol/search/display/view.w3p;page=0;query=Id%3A%22f03d693e-b8dd-4130-a14f-be68ea23198d%22%20Status%3Apublished;rec=0>.

³⁹ Luciana Duranti, "The Impact of Digital Technology on Archival Science," *Archival Science* 1, no. 1 (2001): 49, <http://dx.doi.org/10.1007/BF02435638>.

in outsourcing government records to third party service providers, ensuring that adequate measures are in place for the agency to exercise control over the identity and integrity of its records.⁴⁰

Nevertheless, there are some archival legislation which attempt to specify the roles and responsibilities of both the record creator and preserver. For example, the *Public Records (Scotland) Act 2011* states that government agencies should submit a records management plan to the archival authority.⁴¹ The archival authority, in turn, is required to submit an annual plan to the Scottish Ministers on the records management plans submitted by the agencies. The archival authority needs to include details such as the records management reviews that they conducted, as well as the “names of any authorities that have failed to comply with any of the requirements of an action notice together with details of the alleged failures.”⁴² Another example of an archives act which states the recordkeeping responsibilities of both public offices and the archival authority is the Public Records Act (2005) in New Zealand. The act specifies the recordkeeping responsibilities of public officers with regard to the creation and maintenance of “full and accurate records of its affairs, in accordance with normal, prudence business practice, including the records of any matter that is contracted out to an independent contractor.”⁴³ The act requires that records under the control of public officers must be maintained in an “accessible form” until the disposal of the records are “authorised by or under this Act or required by or under another Act.”⁴⁴ The act also articulates the role of the Chief Archivist in issuing standards relating to the creation, maintenance, appraisal and access to records. The Chief Archivist is expected to present a report to the Minister on recordkeeping practices in public officers annually, and to conduct an audit of recordkeeping practices of the public sector.⁴⁵

Specifying the roles and responsibilities of both the record creator and the preserver in the archival legislation is important in a digital environment, as it is based on the premise that “management of digital records must proceed from a comprehensive understanding of all phases or stages in the lifecycle of records, from the time they are generated, through their maintenance by their creator, and during their appraisal, disposition and long-term preservation as authentic memorials of the actions and matters of which they are a part.”⁴⁶ It will also be useful if archival legislation provides a definition of archival concepts like reliability and authenticity, which would apply to both the record creator and the archives. For example, authenticity is the trustworthiness of a record as a record and is dependent on the “record’s state, mode and form of transmission, and to the manner of its preservation and custody.”⁴⁷ As such,

⁴⁰ ICA International Terminology Database, accessed 20 July 2012, <http://www.web-denizen.com/>.

⁴¹ The National Archives of Scotland merged with the General Register Office for Scotland in April 2011 to become an entity known as the National Records of Scotland. *Public Records (Scotland) Act*, 2011, accessed 21 July 2012, <http://www.nas.gov.uk/documents/PublicRecordsScotlandActPublished.pdf>.

⁴² *Public Records (Scotland) Act*, 2011, accessed 21 July 2012, <http://www.nas.gov.uk/documents/PublicRecordsScotlandActPublished.pdf>.

⁴³ *Public Records Act*, *Parliamentary Counsel Office* 2005, no. 14, <http://www.legislation.govt.nz/act/public/2005/0040/latest/DLM345529.html>.

⁴⁴ *Ibid.*

⁴⁵ *Ibid.*

⁴⁶ Yvette Hackett, Domain 3 Task Force, “Appendix 21: Preserver Guidelines – Preserving Digital Records: Guidelines for Organizations,” [electronic version] in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records*, ed. Luciana Duranti and Randy Preston (Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008), 734. http://www.interpares.org/display_file.cfm?doc=ip2_book_appendix_21.pdf.

⁴⁷ Luciana Duranti, “The Reliability and Authenticity of Electronic Records,” in *Preservation of the Integrity of Electronic Records*, ed. Luciana Duranti et al. (The Netherlands: Kluwer Academic Publishers, 2002), 27.

maintaining and preserving the authenticity of records is a joint responsibility between both the record creator and the archives.

The second weakness of the archival legislation in Commonwealth countries is that the acts typically state that government agencies should seek permission from the national archives before destroying public records. Such a clause works on the assumption that appraisal of records takes place only at the end of the lifecycle when the records become inactive. In reality, appraisal of records, particularly in the digital environment, should be conducted during the active stage of the record's lifecycle. This will enable the archivist to obtain documentation about the recordkeeping environment of the creating agency as well as technical documentation on the digital system which creates and maintains the records.⁴⁸ There is also a need to monitor the appraised electronic records before the records are transferred into archival custody so as to ensure that there are no significant changes in the records and the recordkeeping environment, which can affect the identity and integrity of the records.⁴⁹ Moreover, the timely appraisal of records is in line with the chain of preservation concept. This concept is based on the premise that "from the perspective of long-term or continuing or enduring preservation, all the activities to manage records throughout their existence are linked, as in a chain, and interdependent."⁵⁰ Last but not least, the absence of the use of the term appraisal in a number of archives legislation means that archivists lack "adequate legislative or policy foundations" and that "without a proper mandate to act towards achieving their goal, archivists work at a distinct disadvantage."⁵¹

Another limitation of the archival legislation in Commonwealth countries is they tend to state that records should be transferred to archival custody several decades after they have become inactive.⁵² For example, the Public Record Act (1958) stipulates that public records "shall be transferred not later than thirty years after their creation either to the Public Record Office or to such other place of deposit appointed by the Lord Chancellor under this Act as the Lord Chancellor may direct." Although records may be transferred to archival custody earlier than 30 years between the archives and the transferring agency, the archives act works on the assumption that preservation requirements tend to be incorporated during the end of the record's lifecycle. The lengthy time frame for the transfer of records is not an

⁴⁸ Appraisal Task Force, International Research on Permanent Authentic Records in Electronic Systems (InterPARES), "Appraisal Task Force Report," (2001), p. 19.

http://www.inter pares.org/book/inter pares_book_e_part2.pdf; Yvette Hackett, "Methods of Appraisal and Preservation - Domain 3 Task Force Report," in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2 Experiential, Interactive and Dynamic Records*, ed. Luciana Duranti and Randy Preston (Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008), 190-191.

⁴⁹ Appraisal Task Force, International Research on Permanent Authentic Records in Electronic Systems (InterPARES), "Appraisal Task Force Report," (2001), p. 14.

http://www.inter pares.org/book/inter pares_book_e_part2.pdf

⁵⁰ Terry Eastwood, Randy Preston and Hans Hofman, "Modeling Digital Records Creation, Maintenance and Preservation," in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2 Experiential, Interactive and Dynamic Records*, ed. Luciana Duranti and Randy Preston (Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008), p. 229.

⁵¹ Terry Eastwood, "Reflections on the Goal of Archival Appraisal in Democratic Societies," *Archivaria* 54 (Fall, 2002): 69.

⁵² F. Foscari, "InterPARES 2 and the Records-Related Legislation of the European Union," *Archivaria* 63 (2007), 127, <http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13131/14376>; Chris Hurley, "From Dustbins to Disk-Drives: A Survey of Archives Legislation in Australia," in *The Records Continuum - Ian Maclean and Australian Archives First Fifty Years*, ed. Sue McKemmish and Michael Piggott (Clayton: Ancora Press, 1994), 206-233.

effective strategy to address the management of digital records, where preservation requirements should be “incorporated and manifested in the design of record-making and recordkeeping systems.”⁵³

Finally, archival legislation in Commonwealth countries has varied ways in defining the concept of records and archives. Some of these definitions linked records and archives in terms of its value, in relation to an activity, the passage of time and transfer to an archival institution. For example, the archives act in Singapore differentiates public records as “records of any kind whatsoever produced or received by any public office in the transaction of official business or by any officer in the course of his official duties,” whereas public archives are those records which are “more than 25 years old” of “national or historical significance” and which have been transferred to archival custody.⁵⁴ Such a definition tends to associate public records and public archives in terms of physical placement, which results in a conceptual divide between the management of active and inactive records.⁵⁵ The movement towards increasing privatisation within the public sector means that records from such organizations that do not fall into the schedule of public bodies are potentially excluded from the archives act. Furthermore, such a definition is limiting not only in a traditional analogue environment but also in the digital environment. In the digital environment, it may be potentially contentious to determine which instantiation of the same digital entity belong to whom, especially when there are multiple service providers.

5. Conclusion - Call for Future Empirical Research

The weakness of the archival legislation illustrates the inadequacy, “under-inclusiveness” and the “obsolescence of existing legal rules” in dealing with the challenges posed by the digital environment and the changing nature of public administration.⁵⁶ There is a need to strengthen archival legislation through incorporating archival science including concepts, like the chain of preservation, reliability, and authenticity, as well as the theory of provenance. Duranti notes that the archival legislation of ancient Rome was anchored on the principles of archival science. For example, concepts such as the unbroken chain of custody and the 1898 Dutch manual on arrangement and description were based on early Roman law.⁵⁷ However, with the passage of time, archival legislation departed from the principles and concepts of archival science. As such, there is a need for archivists to revisit these principles and concepts and to work with legislators to strengthen the legislative framework of the archives law.

Finally, one area that requires further empirical research is to understand the perspectives of archivists and records professionals on whether the archives act provides them with a framework to

⁵³ InterPARES 2, “Appendix 19 - A Framework of Principles for the Development of Policies, Strategies and Standards for the Long-term Preservation of Digital Records,” in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2 Experiential, Interactive and Dynamic Records*, ed. Ludiana Duranti and Randy Preston (Padova, Italy Associazione Nazionale Archivistica Italiana, 2008), 709.

⁵⁴ *National Heritage Board Act*, Singapore Statutes Online, 1993, no. 13, <http://160.96.185.113/aol/search/display/view.w3p;page=0;query=Id%3A%22f03d693e-b8dd-4130-a14f-be68ea23198d%22%20Status%3Apublished;rec=0>.

⁵⁵ V. Lemieux, “Archival Solitudes: The Impact on Appraisal and Acquisition of Legislative Concepts of Records and Archives,” *Archivaria* 1, no. 35 (1992), 156, <http://journals.sfu.ca/archivar/index.php/archivaria/article/view/11894/12847>.

⁵⁶ F. Foscari, “InterPARES 2 and the Records-Related Legislation of the European Union,” *Archivaria* 63 (2007): 121-136, <http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13131/14376>.

⁵⁷ Luciana Duranti, “Archival Science,” in *Encyclopedia of Library and Information Science*, vol. 59, ed. Allen Kent (New York, Basel, Hong Kong: Marcel Dekker, 1996), 2-5.

effectively manage records. The application and implementation of an archival legislation involves interaction among stakeholders from different government agencies, including the national archives, and this process of interaction is not adequately covered in case law. Another related area of research is to explore organizational dynamics and interactions among archivists and government representatives as they negotiate their responsibilities in relation to an archives act, including how they interpret and apply an archives act in the management of records. For example, Hurley observed that there was the “interplay of bureaucratic politics” in the implementation of the *State Records Act* in New South Wales and the act was even “opposed and watered down by other agencies who believe that their turf is being invaded.”⁵⁸ Hurley’s statement is a reminder to archivists and records professionals that it is useful to go beyond the instrumental approach in understanding law in terms of its “explicit rules, specialised offices and institutions, and determinate procedures.”⁵⁹ There is also a need to understand archival legislation from the bottom-up. Such a perspective will illustrate the gaps between the law in theory and the law and in practice through providing an insight on why and in what circumstances archivists and records professionals interact with the archives act, as well as cases where the archives act is not relied upon by records professionals.

⁵⁸ Chris Hurley, “From Dust Bins to Disk-Drives and Now to Dispersal: The State Records Act 1988 (New South Wales),” *Archives and Manuscripts* 26, no. 2 (November 1998): 393.

⁵⁹ Roderick A. Macdonald, “Legislation and Governance,” in *Rediscovering Fuller: Essays on Implicit Law and Institutional Design*, ed. Willem J. Witteveen and Wilbren Van der Burg (Amsterdam: Amsterdam University Press, 1999), 310.

Digital Curation: Convergence of Challenges, Institutions and Knowledge

Digital Curation

The Challenge Driving Convergence across Memory Institutions

Sarah Higgins

Abstract

Collaboration between libraries, archives and museums (LAMS) is undertaken in a continuum which starts with an initial understanding of the differences between the disciplines, and can lead to full convergence with a shared mission and delivery of shared services. Collaboration brings increasing benefits in resource efficiencies and user uptake as participating organizations progress through the continuum. It is in the area of digital content creation and management that the synergies of the disciplines are most often harnessed through cooperative exploration, coordinated projects and collaborative services. This paper examines and extends the Collaboration Continuum first identified by Soehner¹ and elaborated by Zorich, Gunter, and Erway² through analyses of the existing research into the nature of LAM collaboration, and identification of the core ethical differences which govern seemingly similar agenda. The paper proposes digital curation as the “change agent” which will bring about full convergence between the professions, as they move through the digital content and management continuum.

Author

Sarah Higgins lectures in Archives Administration and Records Management at Aberystwyth University, where her research focuses on the lifecycle management of digital materials. She was formerly at the Digital Curation Centre (DCC) where she led the DCC Curation Lifecycle Model Project and the standards advisory function. She moved to the DCC from the University of Edinburgh where she undertook various metadata development and co-ordination roles across their cultural collections. A trained cartographer her first archival role was curating the British Antarctic Survey’s Geographical Information Collection and acting as Secretary to the UK Government’s Advisory Committee on Antarctic Place-names.

1. Introduction

Libraries, archives and museums (LAM) are increasingly undertaking joint activities to address gaps in their resources or skills, or to create new services for their existing users; and greater visibility to potential users. Such activities, endorsed by their respective professional organisations, embrace: trust building activities such as professional networking events; coordination activities such as co-location of services; and collaborative activities such as joint outreach programmes. However, it is in the area of catalogue federation and digital content creation and management that collaborative projects have started to lead to shared services. As professional best practice develops and projects mature, LAMs are starting to

¹ Ken Soehner, “Out of the Ring And Into the Future: The Power of Collaboration” (paper presented at the RLG Members Forum “Libraries, Archives and Museums—Three Ring Circus, One Big Show?” July 14, 2005), <http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070504/viewer/file1201.doc>.

² Diane Zorich, Günter Waibel, and Ricky Erway, *Beyond The Silos of LAMS: Collaboration Among Libraries, Archives and Museums* (Dublin, OH: OCLC Online Computer Library Center, 2008), <http://www.oclc.org/research/publications/library/2008/2008-05.pdf>.

converge over the need to provide for the long-term access, use and reuse of their digital materials, both digitized and born-digital, through the new discipline of digital curation.³

2. Collaboration across LAMS

LAMs have maintained a distinct identity, and the commonalities of their core remits and procedures “have not historically been dominant features in the self-characterization of libraries, museums and archives.”⁴ However the possibilities afforded by cross-collaboration have been examined for a number of years with a number of high profile conferences, initiated by the library profession, addressing the topic. The Library Automation Group Conference in 2000 (*Archives, Libraries and Museums Convergence*) looked in detail at cooperative digital projects across LAMs.⁵ The *Choices and Challenges Conferences* of 2002 and 2004 identified the dominant themes of “invisibility, advocacy, convergence and collaboration, and focusing on the researchers and visitors” realising that “... preserving the record, whether it is a text, a photograph, or an object requires collaboration among cultural resource professionals, researchers, visitors, and the public.”⁶

In 2005 the Research Libraries Group (RLG) Conference (*Libraries, Archives, and Museums: Three-ring Circus, One Big Show?*) explored the possibilities for federated searching across collections. This was followed in 2006 by the Rare Books and Manuscripts Section (RBMS) of the American Library Association (ALA) Conference (*Libraries, Archives, and Museums in the Twenty-First Century: Intersecting Missions, Converging Futures?*) taking the conversation a step further by examining the question “to what extent do common goals imply common means?” through a series of case study presentations.⁷ The American Society for Information Science & Technology (ASIS&T) 2009 panel session (*Information Organization in Libraries, Archives and Museums: Converging Practices and Collaboration Opportunities*) examined the range of collaboration opportunities across the sectors and the convergence of information organization practices.⁸ The conversation turned global in 2011 with the International Conference on the Convergence of Libraries, Archives and Museums (ICLAM) 2011 (*User Empowerment through Digital Technologies*) scrutinising the advantages of a converged information environment for the user.⁹

³ Sarah Higgins, “Digital Curation: The Emergence of a New Discipline,” *The International Journal of Digital Curation* 6, no. 2 (2011): 78-88, accessed 30 July 2012, <http://www.ijdc.net/index.php/ijdc/article/view/184/251>.

⁴ J. Trant, “Emerging Convergence? Thoughts on Museums, Archives, Libraries and Professional Training,” *Museum Management and Curatorship* 24, no. 4 (2009): 369-386, p. 370.

⁵ Peter Limb, “Archives, Libraries and Museums Convergence: The 24th Library Systems Seminar, Paris 12-14 April 2000=Archives, bibliothèques et musées,” *Online Information Review* 27, no. 5 (2003): 368.

⁶ Elizabeth Yakel, “Choices and Challenges: Cross-Cutting Themes in Archives and Museums,” *OCLC Systems & Services* 21, no. 1 (2005): 13-17, <http://www.emeraldinsight.com/10.1108/10650750510578091>.

⁷ Christian Dupont, “Libraries, Archives and Museums in the Twenty-First Century: Intersecting Missions, Converging Futures?” *RBM; A Journal of Rare Books, Manuscripts and Cultural Heritage* 8 (Spring 2007): 13-19, accessed 30 July 2012, <http://rbm.highwire.org/content/8/1/13.full.pdf>.

⁸ Ingrid Hsieh-Yee et al., “Information Organization in Libraries, Archives and Museums: Converging Practices and Collaboration Opportunities,” in *Proceedings of the American Society for Information Science and Technology*, vol. 5, issue 1, 2009, pp. 1-5, accessed 30 July 2012, <http://www.asis.org/Conferences/AM09/open-proceedings/panels/36.xml>.

⁹ International Conference on the Convergence of Libraries Archives and Museums: User Empowerment through Digital Technologies, February 15-17, 2011, accessed 30 July 2012, http://www.nift.ac.in/ICLAM_2011/index.htm.

Collaboration between LAMs is being encouraged at the highest professional level, with IFLA, ICA and ICOM joining with the International Council on Monuments and Sites (ICOMOS) and the Coordinating Council of Audiovisual Archives Associations (CCAAA) to form the International NGO Working Group on Convergence in 2008.¹⁰ Later renamed the Libraries, Archives, Museums, Monuments and Sites (LAMMS) Coordinating Council, the group's projects focus on the areas of: copyright and other legal matters, working within the World Intellectual Property Organization (WIPO); political lobbying and practical measures to ensure the safety of cultural heritage, within the activities of the Blue Shield and UNESCO; and the development and standardization for global digital libraries.¹¹

Few research projects have investigated the nature of collaborative activity across LAMs. A comparison of museum and library collaborations in England and the USA identified the short-term subject based project nature of these, with the English focusing on history, while the US focused on art and literature.¹² Yeates and Guy exam cross-domain collaboration to develop a local interest federated resource, identifying the driving force of the projects technical staff.¹³ A report funded by IFLA identified, examined and classified a wide range of collaborative activities across archives, museums and public libraries at local level, in the developed world, formulating a best practice guide for their success.¹⁴ Collaborative activity was found to be taking place in: event programming, integrated facilities such as education and exhibition space, and collaborative digital resource creation. An OCLC research project from 2008-2010 built on the RLG and RBMS conferences to examine collaboration in the context of LAMs with the same organizational governance, which were already committed to working together.¹⁵ Their series of five mediated workshops identified a shared vision "of seamless collections access and community engagement on local Web sites."¹⁶ Again collaborative activity was identified as project based, with the participating organizations focusing on shared creation and storage of digital materials, and the provision of search tools for their discovery. They identified that "with the ever increasing acquisition of born-digital materials, traditional boundaries begin to blend."¹⁷ Lee considers how political and cultural barriers in pan-continental projects can compromise the ability to provide consistent

¹⁰ Ingeborg Verheul, "International NGOs Bonding for Convergence of Libraries, Archives and Museums," 2008, accessed 30 July 2012, <http://www.ifla.org/files/lamms/documents/convergence-communicue.pdf>.

¹¹ International Federation of Library Associations, "About LAMMS," *International Federation of Library Associations Website*, 2010, accessed 30 July 2012, <http://www.ifla.org/en/about-lamms>; Nancy E. Gwinn, "LAMMS and International Collaboration," in *ICOMOS Scientific Symposium – Malta, 2009 Changing World, Changing Views of Heritage: The Impact of Global Change on Cultural Heritage – Technological Change*, 2009, accessed 30 July 2012, http://www.icomos.org/adcom/malta2009/pdf/ADCOM_200910_SYMP_1_Documentation_Nancy_Gwinn.pdf.

¹² Hannah Gibson, Anne Morris, and Marigold Cleeve, "Links Between Libraries and Museums: Investigating Museum-Library Collaboration in England and the USA," *Libri* 57, no. 2 (2007): 53-64, accessed 30 July 2012, <http://198.173.123.161/pdf/2007-2pp53-64.pdf>.

¹³ Robin Yeates and Damon Guy, "Collaborative Working for Large Digitisation Projects," *Program: electronic library and information systems* 40, no. 2 (2006): 137-156.

¹⁴ A. Yarrow, B. Clubb, and J. Draper, *Public libraries, archives and museums: Trends in collaboration and cooperation. IFLA Professional Reports No. 108*, 2008, <http://archive.ifla.org/VII/s8/pub/Profrep108.pdf>.

¹⁵ Diane Zorich, Gunter Waibel, and Ricky Erway, *Beyond The Silos of LAMs: Collaboration Among Libraries, Archives and Museums*, 2008, accessed 30 July 2012, <http://www.oclc.org/research/publications/library/2008/2008-05.pdf>.

¹⁶ *Ibid.*, p. 15.

¹⁷ *Ibid.*, p. 14.

commitment for collaborative ventures, identifying successful projects across Europe, North America and East Asia.¹⁸

3. Cumulative Collaboration

Collaboration across similar, but fundamentally different organizations cannot develop without a mutual trust, which needs to be established over time. Zorich, Waibel, and Erway,¹⁹ building on an address from the RLG Forum,²⁰ articulate a collaborative continuum model which identifies the cumulative development of such trust, and the increasingly sophisticated investment, risk and benefits which accrue. The cumulative nature of collaboration has also been noted by Dornseif²¹ who identified different levels of commitment in collaborative activity between co-located services. These two models of cumulative collaboration can be combined with the roles and responsibilities, benefits, risks and measures for success identified by the various research studies discussed above, to build an understanding of how as collaborative activities mature, commitments, benefits and risk increase, while organizations become more inter-dependent (Figure 1).

Collaborative activity between LAMs starts with *conversance*. An understanding of the professional landscape is developed by keeping abreast of professional developments through channels such as the media, newsletters, webmail or RSS feeds. This low risk activity builds an understanding of the possibilities afforded by collaboration. Conversance may lead to *contact* either in person or through social media channels such as Twitter or Facebook. This casual experimental networking enables the exploration of commonalities such as subject, place, audience or aim. Minimal resources are expended and no commitments are made, however collaborative ideas may start to emerge and a trust relationship may start to grow. Minimal integration *cooperative* activities such as meeting together to share information on a shared point of interest will build further trust so that a vision of future activities can start to develop. The risk increases as LAMs move into *coordination*. Short-term jointly coordinated projects need a commitment from all parties to work together with a common goal to undertake an activity for a clearly defined audience. Such projects include Web based or physical joint exhibitions, educational activities or lecture programmes. Success in such selective projects requires guidance from: an engaged management; a dedicated project manager and clear aims and objectives with feasible timelines. Projects need the benefit of a stable administrative base and resources, which have been carefully planned, along with a commitment to timely communication. The risks increase proportionately to the resource expended, and projects can founder on a lack of commitment to overcome differences in procedures, priorities and language. Even one or two staff with a resistant attitude to coordinated activity can risk its success through lack of commitment to the project's goal. Co-location of services is a form of cooperation where a management commitment has risked locating seemingly similar organizations in the same managerial, and sometimes also physical space, so that there is opportunity is created for resources

¹⁸ Hyuk-Jin Lee, "Collaboration in Cultural Heritage Digitisation in East Asia," *Program: electronic library and information systems* 44, no. 4 (2010): 357-373, <http://www.emeraldinsight.com/10.1108/00330331011083248>.

¹⁹ Zorich et al., *Beyond The Silos of the LAMs*, pp. 8-12.

²⁰ Ken Soehner, "Out of the Ring And Into the Future: The Power of Collaboration" (paper presented at Libraries, Archives, & Museums—Three-Ring Circus, One Big Show?, St. Paul, MN, July 14, 2005), accessed 30 July 2012, <http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070504/viewer/file1201.doc>.

²¹ Karen A. Dornseif, "Joint Use Libraries: Balancing Autonomy and Cooperation," *Resource Sharing & Information Networks* 15, no. 1/2 (2001): 103-116.

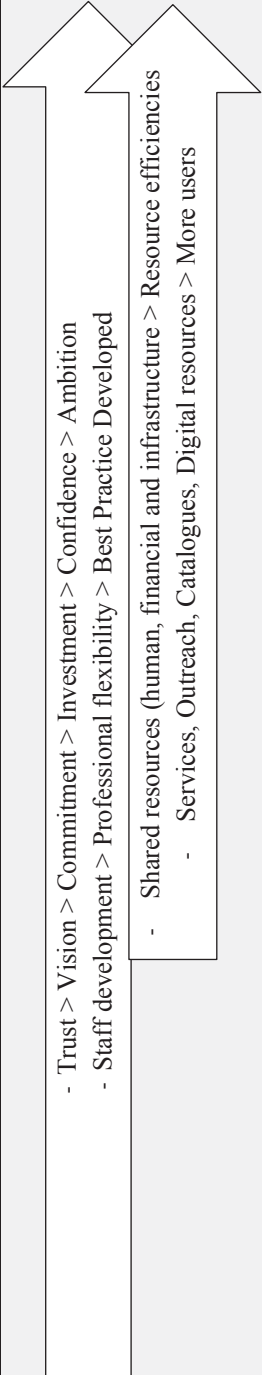
Collaboration Continuum	Conversance	Contact	Cooperation	Coordination	Collaboration	Convergence
Level of Integration	None	Experimental	Minimal	Selective		Full
Commitment	None	Casual networking	Trust building	Shared goal with a separable administrative framework	Shared vision with changed working practices and inter-dependencies	Shared mission and service delivery
Activities	Keeping abreast of professional developments	Exploration of differences and commonalities	Information sharing	Joint projects	Joint projects leading to shared services	Shared infrastructure
Triggers for Moving Along the Continuum	- Resources gap: skills, financial, infrastructure - Change agent: Management mandate, Staff incentives, Funding availability					
Benefits	<div>  <p> - Trust > Vision > Commitment > Investment > Confidence > Ambition - Staff development > Professional flexibility > Best Practice Developed - Shared resources (human, financial and infrastructure > Resource efficiencies - Services, Outreach, Catalogues, Digital resources > More users </p> </div>					
Risks	Minimal	Minimal <ul style="list-style-type: none"> - Low resource commitment - Easy to walk away 	<ul style="list-style-type: none"> - Lack of common goal /vision and resistant attitudes - Different procedures and priorities - Lack of common language - Insufficient resource capacity - Domination by one organization - Problems with intellectual property rights - Management commitment and involvement - Engaged and committed staff (through incentives or encouragement) - Tangible stable resources (human, financial and infrastructure) - Clear aims and objectives - Clearly defined audience - Staff training 			<ul style="list-style-type: none"> - Differences not addressed prior to commitment
Measures for success	Knowledge of possible collaborators	<ul style="list-style-type: none"> - Trust develops - Ideas for joint working emerge 		<ul style="list-style-type: none"> - Dedicated project co-ordinator - Administrative project base - Feasible timelines - Timely communication 		
						Harmonization of: <ul style="list-style-type: none"> - Organizational procedures - Collection Management - Framework - Priorities

Figure 1. The Extended Collaboration Continuum.

to be shared and collaborations to develop “co-location of the museum and library might make collaboration more likely to occur and easier when it happened.”¹ *Collaboration* deepens the commitment between LAMs, so that in certain areas they become inter-dependent and working practices have to change. A shared vision is vital for substantive joint projects such as a shared IT infrastructure, shared provision of services, and this may be missing in forced colocation risking the success of the venture. Collaboration carries a greater risk than coordination and may require formal memorandums of agreement to ensure management commitment and resource allocation. *Convergence* happens when joint activities cease to be selective collaborations which require special project status. Instead there is a full commitment to integration around a common mission. Organizational procedures, the collection management framework and priorities are harmonized so that shared services are delivered from a shared infrastructure. There are risks associated with a lack of engagement from management and staff, a lack of resources and a lack of clarity about the aims and objectives of converging, but the main risk is that differences between LAMs are not properly addressed and procedures developed which are acceptable to all parties.

Benefits accrue as you move along the continuum. As trust develops between LAMs, so does the vision for a shared future and the commitment to making it happen. Investment in shared projects reaps results and increases confidence, so that the ambition for the scale and inter-dependence of future projects grows. Development in staff through training and increased experience leads to professional flexibility and the incremental development of best practice. Additional benefits accrue from coordination, collaboration and convergence. Shared resources, be they human, financial or through shared infrastructure can lead to efficiencies. Such efficiencies can enable better services, invigorated outreach programmes, improved catalogues and the creation of digital resources. The result is greater visibility and increased user numbers.

A number of triggers prompt movement along the collaboration continuum. The identification of a resource gap may start an investigation into possible partners to help plug the gap, and maybe partner in a bid for project funding. Gaps may relate to financial constraints for activities outside the core service, a lack of applicable skills for new initiatives, to develop infrastructures for new modes of service or IT, or a lack of suitable space. Moves along the continuum are generally started by what Zorich, Waibel, and Erway termed a “... ‘change agent’—a trusted individual, partner or program that keeps the effort alive, injects it with a dose of resources (ideas, technology, staff) at the right time and keeps the participants focused on the overall vision they are aiming to bring to life.”²

Change agents are characterized by Zorich, Waibel, and Erway as a “trusted individual, partner or program” a LAM,³ but they are as likely to be a co-ordinator appointed to keep project on track, to be an umbrella policy organisation which fosters a spirit of collaboration, or one which makes available an appropriate funding stream. Collaborative activities championed or funded by a *change agent* need the support of management through a mandate if a shared vision is to develop. Staff incentives, both financial or through professional recognition, such as promotion or the chance to present their work at meetings and conferences can help to keep a vision alive.

¹ Gibson, Morris, and Cleeve, “Investigating Museum-Library Collaboration in England and the USA.”

² Zorich et al., *Beyond the Silos of the LAMs*, p. 24.

³ Ibid.

4. Conversance and Contact

The distinctions between LAMs are deep rooted in the separate identities of the professions, and before any progress can be made along the collaboration continuum, LAM professionals need *conversance* of these differences, so that trust can start to develop in a spirit of mutual understanding and respect.

The international bodies representing the three professions: the International Council on Archives (ICA), the International Council of Museums (ICOM) and the International Federation of Library Associations (IFLA) each publish a code of ethics (the latter in draft at the time of writing) which establish the values, principles and activities for the respective professions. These show a marked difference in their content, and core divergences in the emphases of the three professions.

*The ICA Code of Ethics*⁴ stresses the management of the materials in their care, defined as the documentary heritage: records, documents and archival material. The emphasis is on maintaining the characteristics of an authoritative record, which are defined by *ISO 15489: The International Standard for Records Management* as: authenticity, reliability, integrity and usability.⁵ The individual practitioner's primary duty is identified as the maintenance of the integrity of the records to ensure their reliable evidential nature. Protection of the authenticity of records is necessary to maintain the archival value of records, while ensuring usability through maintaining continuing access and intelligibility. The Code stresses the implementation of "good recordkeeping practices throughout the life-cycle of documents"⁶ to maintain "the archival value of records."⁷ Archival value is maintained through practicing archival principles when undertaking and recording the recordkeeping tasks of: appraisal, arrangement, description, preservation and conservation. Archival principles, first defined in 1898⁸ still form the basis of professional activity.

- Principle of provenance: "records created or received by one recordskeeping unit should not be intermixed with those of any other."⁹
- Principle of original order: "records should be maintained in the order in which they were placed by the organization, individual, or family that created them."¹⁰

For archivists the initial selection of material should be undertaken with impartiality and due regard the evidential value, while taking account of corporate and personal privacy. Providing access to the materials is not the primary concern of *The ICA Code*, and although the provision of an archival service is expected, there are a number of caveats concerning possible imposition of access restrictions.

⁴ International Council on Archives, *International Council on Archives Code of Ethics*, 1996, accessed 30 July 2012, <http://www.ica.org/download.php?id=574>.

⁵ International Organization for Standardization, *ISO 15489-1:2001, Information and Documentation - Records Management - Part 1: General*, 2001.

⁶ International Council on Archives, *International Council on Archives: Code of Ethics*, 1996, Clause 5, <http://www.wien2004.ica.org/sites/default/files/Ethics-EN.pdf>.

⁷ *Ibid.*, Clause 3.

⁸ Samuel Muller, Johan Feith, and Robert Fruin, *Manual for the arrangement and description of archives drawn up by the direction of the Netherlands Association of Archivists*, 2nd ed. (New York: Wilson, 1968).

⁹ Maygene Daniels, "Introduction to Archival Terminology," in *Web Version based on: A Modern Archives Reader: Basic Readings on Archival Theory and Practice*, Web edition. (National Archives Trust Fund Board, 1999), accessed 30 July 2012, <http://www.archives.gov/research/alic/reference/archives-resources/terminology.html>.

¹⁰ *Ibid.*

By contrast the draft *IFLA Code of Ethics for Librarians and Other Information Workers* identifies the provision of access to information as the core mission for information professionals.¹¹ Five of its six clauses discuss the professional requirement to provide unrestricted, effective and indiscriminate access to information, with articulation of the profession's support for open source environments and intellectual property rights that enhance creativity. The emphasis on access is linked to societal need to share information and the rights set out in Article 19 of the *Universal Declaration on Human Rights* "to seek, receive and impart information and ideas through any media and regardless of frontiers."¹² However, the code also reflects the *Five Laws of Library Science*¹³ that established the core ethical values of the library profession.

1. Books are for use rather than for preservation – leading to vitalisation of the library
2. Books are for all rather than the chosen few – making the library a responsibility of society for the education of all
3. Books should be openly accessible, arranged and catalogued to promote discovery – enabling both structured and serendipitous discovery
4. Save the time of the reader – library management should be reader-centric enabling fast and efficient retrieval
5. The library is a growing organism – enabling growth and change in the service

Unlike the *ICA Code*, the *IFLA Code* does not dwell on the actual activities required to manage the materials beyond: the need to "organize and present content" to allow autonomous access; and the definition of published policies for "selection, organization, preservation, provision and dissemination of information."

The ICOM Code of Ethics for Museums takes a corporate rather than a personal view of the profession, and as such details the policies, processes and responsibilities required to underpin their responsibility for the natural and cultural heritage, and the legal framework in which museums operate.¹⁴ More holistic than either the *ICA Code* or the *IFLA Code* it identifies both the management of the materials in their care and as the promotion of their collections as the primary responsibilities. The underlying ethos is "that of service to society, the community, the public and its various constituencies, as well as the professionalism of museum practitioners."¹⁵ The current code grew out of the *Ethics of Acquisition* that identified "certain principles of ethics and professional integrity in relation to acquisition" and reinforced the ethical link between acquisition of materials and their use for research,

¹¹ International Federation of Library Associations, (*International / IFLA-) Code of Ethics for Librarians and Other Information Workers, Draft* (The Hague, 2011), accessed 30 July 2012, <http://www.ifla.org/files/faife/news/ICoE-Draft-111208.pdf>.

¹² United Nations, "The Universal Declaration on Human Rights," 1948, accessed 30 July 2012, <http://www.un.org/en/documents/udhr/index.shtml>.

¹³ Shiyali Ramamrita Ranganathan, *The Five Laws of Library Science* (Madras Library Association, 1931), <http://babel.hathitrust.org/cgi/pt?id=uc1.b99721;page=root;view=1up;size=100;seq=9;orient=0>.

¹⁴ International Council of Museums, *ICOM Code of Ethics for Museums* (Paris, October 2006), accessed 30 July 2012, http://icom.museum/fileadmin/user_upload/pdf/Codes/code2006_eng.pdf.

¹⁵ *Ibid.*, p. iv.

education and conservation.¹⁶ This link between collecting and service to a user community was established in the first *Code of Ethics for Museum Workers*:¹⁷ “their [museums] value is in direct proportion to the service they render the emotional and intellectual life of the people. The life of a museum worker is essentially one of service.”¹⁸

The rounded inter-relationships of ethical concerns for the museum profession have been summarized as an equal balance between collections care and public service.¹⁹ This balanced ethical outlook contrasts with the collection centric view of the archive profession and the user centric view of the library profession. The fundamental differences in character between archives and museums have been summarized as “archives are more often wholesalers and largely contribute to the products of others

...museums were retailers with direct products for users.”²⁰ Libraries similarly offer a direct product and access methods for their users.

Becoming *conversant* with the practices of a different domain so that the professional landscape can be fully understood, and making *contact* through, for instance, the sort of casual networking offered by the cross-domain conferences highlighted above or social networking channels, can ensure a deeper understanding of the different professional stances in an experimental, low risk and low commitment environment, before any commitments to further activity are made.

5. Cooperation and Coordination

Despite the different collection focuses and divergent ethical emphases, the three *Codes* show that there is cross-sectoral agreement that their function is the care of collections and provision of access to these, so that they can be used. These “key commonalities of collecting” are the basis of the movement along the continuum to *cooperative* and *coordinated* activities.

LAMs are distinguished by institutional persistence and user trust, which are the basis of the development and continuation of knowledge communities.²¹ Crucially, although divergent in the types of material they collect, LAMs all provide sustainable collections which are managed by experts. Their role as “sustained institutions to collect, organize, preserve, and provide access to knowledge-bearing objects” co-evolved their expertise, methodologies and tools for organizing and interpreting knowledge²² so that parallel systems were developed for the core tasks this involves.

¹⁶ International Council of Museums, “Ethics of Acquisition,” 1970, accessed 30 July 2012, <http://www.museum.or.jp/icom-J/acquisition.html>.

¹⁷ Alissandra Cummins, “ICOM’s Ethical Principles,” *Danske Museer* 23, no. 4 (2010), accessed 30 July 2012, http://icomdanmark.dk/etikseminar2010/Cummins_ICOM_ethical_principles_2010.pdf.

¹⁸ American Association of Museums, *Code of Ethics for Museum Workers* (American Association of Museums, 1925).

¹⁹ Gary Edson, “Ethics,” in *Museum Ethics*, ed. Gary Edson (New York: Routledge, 1997), 2-15.

²⁰ Elizabeth Yakel, “Choices and Challenges: Cross-Cutting Themes in Archives and Museums,” *OCLC Systems & Services* 21, no. 1 (2005): 13-17, p. 16, accessed 30 July 2012, <http://www.emeraldinsight.com/10.1108/10650750510578091>.

²¹ Margaret Hedstrom and John King, *On the LAM: Library, Archive, and Museum Collections in the Creation and Maintenance of Knowledge Communities* (Paris, France: Organisation for Economic Co-operation and Development, 2002), accessed 30 July 2012, <http://www.oecd.org/dataoecd/59/63/32126054.pdf>.

²² Ibid.

LAMs can only ensure the sustainability of collections and expertise knowledge management function in an organizational environment which ensures the “physical and moral defence” of their collections.²³ That is the physical and intellectual security of collections, through effective collections management underpinned by policies which address both the organizational environment and the day-to-day administration.

The UK’s *Publicly Available Specification (PAS) 197* identifies the similarities of processes and procedures across the LAMs, articulating a cross-sectoral *Code of Practice for Cultural Collections Management*.²⁴ A thesaurus to the collection processing workflow across the disciplines is provided,²⁵ clearly identifying a limited distinction in information organization practices (Table 1), despite the semantic differences. All three disciplines: identify items to collect; receive them into their care; classify, catalogue and index; and periodically review and remove items from the collection.

Table 1. Collection Processing Workflow in PAS 197.²⁶

Workflow Order	Library Item	Archive Item	Museum Item
1	Selection	Pre-accession	Entry
2	Acquisition	Accession	Acquisition
3	n/a	n/a	Accession
4	n/a	Appraisal	n/a
5	Cataloguing	Cataloguing	Cataloguing
6	De-accession / Withdrawal	De-accession	De-accession / Disposal

The PAS provides a discipline neutral Collections Management Framework which includes these information organization procedures as part of an overall collection management policy, administered within an encompassing organizational remit in which the policies, processes and procedures are constantly monitored through internal audit and management review, so that improvements can be made. The collection management policy also includes the collection administration areas of: collections development, collections access and collections care. A granular list of processes, again domain neutral, associated with each of these areas is supplied in the PAS. The exhaustive list of activities will be familiar to all in the LAM professions and includes, for example, condition checking, displaying and exhibiting, assessing impact and packing.

These similarities can form the basis of cooperative and coordinated activities where areas of synergy can be explored, information on best practice shared and the possibilities of co-location of services can be explored for the benefits of both staff and users, along with financial and space saving benefits.

²³ Hilary Jenkinson, *Manual of Archive Administration*, 2nd ed. (London: Lund Humphries, 1966).

²⁴ British Standards Institution, *PAS 197: Code of Practice for Cultural Collections Management*, 2009.

²⁵ Ibid., p. 6.

²⁶ Ibid.

6. Collaboration for Digital Collections Creation and Management

In the digital age access to documentary heritage is no longer geographically restricted to those who can present themselves in person, relying on a cultural organization's ability to produce a physical object on request. Archival documents and copies of books are increasingly provided in digital format through the Internet, making digital surrogates, cheap to the end user, ubiquitous across information management disciplines. These preserve the original from over-use, but more pertinently provide both remote and multiple user access. Archives and museums are challenged by the management of both digitized and born-digital documents; libraries are dealing with the rise of the eBook; and the academic sector is grappling with institutional repositories and research data management. Collection development is also no longer restricted by the physical location of the material, or guardianship of the trusted organization charged with the primary care of the authentic copy. Collaborative virtual libraries, which serve a community of interest, or bring together related or separated materials, can be readily developed.

Since the late 1990s much collaborative effort has been dedicated to enabling the online delivery, search and discovery of materials held by LAMs. Gibson, Morris, and Cleeve noted that many of the collaborative projects they studied had some level of digitization, federated search, or shared database activity.²⁷ Yarrow, Clubb, and Draper noted global, continental, national and local initiatives to create collaborative digital resources.²⁸ Most strikingly, participants in Zorich, Waibel, and Erway's 2008 study only identified digital initiatives as a basis for their collaborations.²⁹

In the late 1990s the US Digital Libraries Initiative acted as the "change agent" in USA, providing funding to experiment with the possibilities of new technology. In the UK this role was taken through funders such as the Joint Information Systems Committee (JISC), initially through its explorative eLib Programme, which emphasized the practical collaborative application of technologies to providing materials for learning and teaching.³⁰ Other UK funders such as the Arts and Humanities Research Board (AHRC) also provided competitive funding for collaborative digitization projects focusing on providing scholarly materials for research analysis through the new discipline, digital humanities.³¹

A skill base in digitization was initially built through hand-crafted *boutique* digitization projects, "for which the principal goal [was] experimentation with new technologies and extraordinary attention to the unique properties of each artefact."³² As more sophisticated practice developed the focus shifted to "the creation of useful and relevant collections that served the needs of one or more communities of users."³³ Many organizations now undertake mass digitization, or orchestrate the federation or

²⁷ Gibson et al., "Links Between Libraries and Museums."

²⁸ Alexandra Yarrow, Barbara Clubb, and Jennifer-Lynn Draper, *Public Libraries, Archives and Museums: Trends in Collaboration and Cooperation* (The Hague: International Federation of Library Associations and Institutions, 2008), <http://www.ifla.org/files/public-libraries/publications/prof-report-108/108-en.pdf>.

²⁹ Zorich et al., *Beyond The Silos of the LAMs*.

³⁰ Chris Rusbridge, "Towards a Hybrid Library," *D-Lib Magazine* (July/August 1998), <http://dlib.ukoln.ac.uk/dlib/july98/rusbridge/07rusbridge.html> (accessed 30 July 2012).

³¹ King's College London, "Digital Humanities Modules: Why Should You Take Digital Humanities Modules as Part of Your MA?," n.d., accessed 31 July 2012, <http://www.kcl.ac.uk/artshums/depts/history/modules/level7/digi.aspx>.

³² Paul Conway, "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas," *The Library Quarterly* 80, no. 1 (2010): 76.

³³ National Information Standards Organization (NISO), *A framework of guidance for building good digital collections*, 3rd edition (Baltimore, MD, 2007), p. 1, accessed 31 July 2012, <http://www.niso.org/publications/rp/framework3.pdf>.

aggregation of diverse digital materials from a range of cultural organizations. There are many such projects worldwide and only three, with local resonance for the author are mentioned here as examples of the range of activities being undertaken. Europeana is a European scale project originally funded by the European Commission's eContent^{plus} Programme.³⁴ It "...provides single access point to millions of books, paintings, films, museum objects and archival records that have been digitized throughout Europe³⁵ ... [which] enables people to explore the digital resources of Europe's museums, libraries, archives and audio-visual collections"³⁶ through aggregation of resources and metadata cross-walking. The best practice guidelines from the associated Europeana Travel Project³⁷ identify the most appropriate guidance available in the areas of: content creation and access, digitization project management, and user studies.³⁸ On a national level, the People's Collection of Wales³⁹ is a project sponsored by the Welsh Government. It not only aggregates resources from Welsh LAMs, and associated cultural heritage organisations, but taps into the collaborative content creation spirit of Web 2.0, by enabling any group or individuals to participate by contributing their own materials, which then go through a quality assurance process. This project directly addresses the "... tension between the vision of seamless collections access and community engagement on local Web sites, and the shift to online user behaviour where access and engagement now occur at a broader network level" expressed by participants in Zorich, Waibel and Erway's study.⁴⁰

On an individual organizational level, the National Library of Wales undertakes large digitization projects of diverse materials of cultural significance to Wales, as a core funded activity, while leading collaborations in externally funded aggregation projects. A current project, funded by JISC "will conduct mass digitization of primary sources relating to World War One from the Libraries, Special Collections and Archives of Wales."⁴¹

7. Digital Collaboration and Technical Skills

Building digital collections is now a mature activity and holistic international best practice, which addresses all of the socio-economical, organizational and technical challenges, has been developed iteratively. Best practice in digitization, and caring for born-digital materials, has been collated and codified by the US National Information Standards Organisation (NISO); a US led international

³⁴ The eContent^{plus} programme closed on 31 December 2008. Its successor is the Information and Communications Technologies Policy Support Programme which now has Europeana as its first objective – see accessed 31 July 2012, http://ec.europa.eu/information_society/activities/econtentplus/index_en.htm.

³⁵ Europeana Professional, "About Us," *Europeana*, n.d., accessed 31 July 2012, <http://pro.europeana.eu/web/guest/about>.

³⁶ Europeana Think Culture, "Europeana: Think Culture," n.d., accessed 31 July 2012, <http://www.europeana.eu/portal/>.

³⁷ The Europeana Travel Project is now completed but information on it can be found at: <http://www.europeanatravel.eu/> (accessed 31 July 2012).

³⁸ Karmen Štular Sotošek, *Best practice examples in library digitisation* (Europeana Travel, 2011), accessed 30 July 2012, http://www.europeanatravel.eu/downloads/ETravelD2_2final.pdf.

³⁹ People's Collection Wales, "People's Collection Wales," n.d., accessed 31 July 2012, <http://www.peoplescollectionwales.co.uk/>.

⁴⁰ Zorich et al., *Beyond The Silos of LAMs*, p. 15.

⁴¹ Joint Information Systems Committee, "JISC: Strand B: Mass Digitisation," n.d., accessed 31 July 2012, http://www.jisc.ac.uk/whatwedo/programmes/digitisation/content2011_2013/Strand B.aspx.

collaborative effort, intended for the domain neutral “cultural heritage” and “funding” organizations.⁴² This guidance considers *good* digital collections in the context of four sets of principles relating to: 1) collection development and processing; 2) digital object creation and management; 3) metadata creation and management; and 4) public service and project management.

The *initiative principles* of the document, which mostly relate to public service delivery will be familiar activities to LAM professionals concentrating on project management, staffing marketing and evaluation. Meanwhile, eight of the *collection principles*, on face value are activities for which form the basis of LAM professionals ethical codes and for which they are trained. Table 2 shows how they can be mapped directly onto the collection care areas covered by the PAS Collection Management Framework.⁴³

Table 2. Elements of the PAS 197 Collections Management Framework
Mapped Against the NISO Collections Principles.⁴⁴

PAS Collection Management Framework	Collections development policy, processes and procedures	Collections information policy, processes and procedures	Collections access policy, processes and procedures	Collections care and conservation policy, processes and procedures
NISO Collections Principles	Principle 1: A good digital collection is created according to an explicit collection development policy.	Principle 2: Collections should be described so that a user can discover characteristics of the collection, including scope, format, restrictions on access, ownership, and any information significant for determining the collection's authenticity, integrity, and interpretation. Principle 5: A good collection respects intellectual property rights.	Principle 4: A good collection is broadly available and avoids unnecessary impediments to use. Collections should be accessible to persons with disabilities, and usable effectively in conjunction with adaptive technologies. Principle 6: A good collection has mechanisms to supply usage data and other data that allows standardized measures of usefulness to be recorded.	Principle 3: A good collection is curated, which is to say, its resources are actively managed during their entire lifecycle. Principle 9: A good collection is sustainable over time.

However, the actual activities they describe require, in the main, technical or specialist training and knowledge for implementation: access restriction needs knowledge of digital authentication procedures and digital assurance; respecting intellectual property rights needs knowledge of digital governance; and lifecycle management needs knowledge of digital preservation procedures. The remainder of the collection development principles deal with a completely new skill-set which requires IT competency: the

⁴² Ibid., p. 1.

⁴³ British Standards Institution, *PAS 197: Code of Practice for Cultural Collections Management*, p. 7.

⁴⁴ Ibid.; NISO, *A framework of guidance for building good digital collections*.

need for a collection to be interoperable for repurposing and contextual understanding; the need for a collection access methodology to integrate with a users normal workflow through the provision of organizational and collaboration tools; and the need for a collection to be sustainable over time. The same is true for the principles relating to the creation and management of a digital objects. Although the need to assign a persistent identifier and associated metadata, while maintaining contextual information and authenticity are familiar to all LAM professionals, the technical requirements of undertaking these tasks in the digital realm may not be. Exhaustive instructions for the creation of standards compliant, interoperable metadata which uses authority control, content standards and crosslinks to other resources to facilitate cross-searching while supporting the “long-term curation and preservation of objects in collections,”⁴⁵ signposts the need for specialist metadata practitioners to design data models before any cataloguing activity can be undertaken.

Developing and managing collaborative digital collections is now as much a technical endeavour as an information management one, and the emphasis is on the management and delivery of digital collections that “should be accessible through the Web, using technologies that are well known among the target user community,”⁴⁶ while advocating “the entire lifecycle of the digital collection and associated services.”⁴⁷ Successful delivery relies on both information professionals and IT professionals to build an interoperable technical architecture. The ambition, scale and complexity of most endeavours meaning it is no longer possible for a LAM professional with some technical know-how to *have a go*.

7.1 Convergence Through Digital Curation

The NISO guidance articulates an interoperable technical infrastructure, which considers the possibilities of long-term access, use and reuse from the outset, rather than an after-thought when the funding has ended.⁴⁸ Funding agencies are starting to mandate long-term maintenance of digital material created with their funding, in the UK taking their lead from the Research Councils UK’s *Common Principles on Data Policy*.⁴⁹ This focus on “ensuring that digital material is managed throughout its lifecycle so that it remains accessible to those who need to use it”⁵⁰ is the basis of *digital curation*. This new discipline considers all the activities required to maintain access to digital materials over the long-term, rather than the short-term goals of content creation. Long-term management cannot be undertaken as short-term projects, but requires a consistent funding stream, management structure and technical methodologies.

The Open Archival Information Systems Reference Model (OAIS)⁵¹ was identified, as “a common framework for digital preservation applications”⁵² while it was in development. It describes both the information architecture and the organizational requirements for the long-term care of digital material. This has formed the basis of much collaborative activity to provide technical applications including the development of storage solutions in the form of *open repositories*, and metadata standards for both

⁴⁵ Ibid., p. 62.

⁴⁶ Ibid., p. 11.

⁴⁷ Ibid., p. 86.

⁴⁸ Ibid., p. 1.

⁴⁹ Research Councils UK, “Excellence with impact: RCUK common principles on data policy,” n.d., <http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx>.

⁵⁰ Higgins, “Digital Curation,” p. 79.

⁵¹ International Organization for Standardization, *ISO 14721: Space data and information transfer systems — Open archival information systems — Reference model*, 2003.

⁵² Higgins, “Digital Curation,” p. 80.

creating information packages⁵³ and describing preservation activities.⁵⁴ Like digital content creation, many of the activities required form the basis of LAM ethical codes. The *mandatory responsibilities* described by OAIS⁵⁵ can be mapped directly to the core activities of information professionals described in both their ethical codes and PAS 197 (Table 3). However, the technical challenge requires co-working with IT professionals, and specialist training to enable this are seeing the development of converged services.

Table 3. The Mandatory Responsibilities of OAIS Mapped Against Collection Management Procedures for LAMs.⁵⁶

Mandatory Responsibility of OAIS	Information Professionals Core Activities
Collect archival materials from the creator and accession into the repository supported by Collections Policy, Acquisition Policy and Accessions Policy.	Collect materials and accession into the collection supported by Collections Policy, Acquisition Policy and Accessions Policy.
Arrange, describe and ensure finding-aids are available for the material. Prepare the material for storage by removing anything harmful to long-term preservation, packaging appropriately and store in a suitable environment.	Arrange, describe and catalogue the material. Prepare the material for storage and store in a suitable environment.
Develop an Access Policy and access methodology to ensure the material can be made available to the identified users.	Develop an Access Policy and access methodology to ensure the material can be made available to the identified users.
Provide contextual information through arrangement and description—catalogues, finding-aids and interpretive materials.	Provide contextual information through provision of catalogues, finding-aids and interpretive materials.
Implement a Preservation Policy which ensures the materials do not deteriorate and are handled appropriately. Ensure secure storage so that records are not tampered with or inappropriately copied.	Implement a Preservation Policy which ensures the materials do not deteriorate, are handled appropriately and securely stored.
Ensure provision and procedures for access are in place for the identified users.	Ensure provision and procedures for access are in place for the identified users.

Digital curation is the *change agent* which is moving LAMs along the collaboration continuum from *collaboration* to *convergence*. The technical challenges and investment required mean that organizations are forced to pool experience and expertise. Best practice and technical methodologies are developing

⁵³ The Metadata Encoding Transmission Standard (METS) was developed explicitly to address the Information Model identified in section 4.2 of OAIS. More information can be found at: <http://www.loc.gov/standards/mets/> (accessed 31 July 2012).

⁵⁴ The PREMIS Data Dictionary for Preservation Metadata took OAIS as a developmental start point. The full standard can be found at: <http://www.loc.gov/standards/premis/v2/premis-2-2.pdf> (accessed 31 July 2012).

⁵⁵ International Organization for Standardization, ISO 14721: Space data and information transfer systems — Open archival information systems — Reference model. p. 3-1.

⁵⁶ Ibid; Sarah Higgins, “Digital Curation: New Medium, Old Methods” (paper presented at Advocating for Archives and Records: The Impact of the Profession in the 21st Century, Edinburgh, 31 August - 2 September 2011).

through consortia discussion and training⁵⁷ and collaborations developing so that shared missions and shared services and infrastructure are starting to emerge such as the MetaArchive Cooperative in the US, which openly declares interdependence between partners.⁵⁸ An example, local to the author, is the development of a shared digital curation service across Wales, being led by the Archives and Records Council Wales (ARCW).

Digital curation is now articulating its own agenda, based around this convergence of the core skills of LAM professionals and the discipline of informatics. Domain neutral continuing professional development courses, and awareness raising events, to bring LAM professionals up to speed; and optional training as part of relevant university courses, are now maturing into dedicated degree programmes that address both the procedural and technical skills required for information engineering.⁵⁹ Ataman has identified mainly technical skills in his list of *add-ons* for the information training: content presentation and management; trustworthy storage, digital forensics and e-curation, giving records management the procedural role.⁶⁰ Simmons College in the US received a grant in 2009 “for the development of a Cultural Heritage Informatics curriculum specifically designed to address the digital convergence of cultural heritage institutions - libraries, archives, and museums,”⁶¹ which brings together the social, technical, cultural and political concerns for stewardship to ensure *curation ready data*.⁶² Lee and Tibbo during the DigCCurr project analysed the core knowledge and capabilities which need to be developed, and the rationale behind these, to develop a holistic curriculum that considers both the conceptual and practical skills required by digital curation professionals. They identified the need to teach a professional context for digital curation with values and principles as a discipline in its own right, along with the technical requirements for ensuring digital material remains useable and useful.⁶³

9. Digital Creation and Management Continuum

Digital creation and management has followed its own continuum becoming increasingly sophisticated as technologies advance, the possibilities for their implementation are explored, and best practice develops.

⁵⁷ In the UK the Digital Preservation Coalition (DPC), a cross-domain collaborative membership organization offers training, support and networking. The Digital Curation Centre (DCC) also offers these, but focuses on research data management in Higher Education.

⁵⁸ Educopia Institute, “MetaArchive Cooperative: Declaration of Interdependence,” 2012, <http://www.metaarchive.org/declaration>.

⁵⁹ Christopher A. Lee and Helen Tibbo, “Where’s the Archivist in Digital Curation? Exploring the Possibilities through a Matrix of Knowledge and Skills,” *Archivaria* 72(Fall 2011): 123-168, <http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13362/14668>.

⁶⁰ Bekir Kemal Ataman, “Requirements For Information Professionals in a Digital Environment: Some Thoughts,” *Program: Electronic Library and Information Systems* 43, no. 2 (2009): 215-228, <http://www.emeraldinsight.com/10.1108/00330330910954415>.

⁶¹ Jeannette A. Bastian, Michele V. Cloonan, and Ross Harvey, “From Teacher to Learner to User: Developing a Digital Stewardship Pedagogy,” *Library Trends* 59, no. 4 (2011): 607-622, accessed 31 July 2012, http://muse.jhu.edu/content/crossref/journals/library_trends/v059/59.4.bastian.html

⁶² The properties of curation ready data can be found in Figure 2.2 of: Sarah Higgins, “The lifecycle of data management,” in *Managing Research Data*, ed. Graham Pryor (Facet Publishing, 2012).

⁶³ Helen Tibbo and Christopher Lee, “Convergence through Capabilities: Digital Curation Education for Libraries, Archives and Museums,” in *Archiving 2010 (IS&T, 2010)*, pp. 53-57, accessed 31 July 2012, <http://www.ils.unc.edu/callee/p53-tibbo.pdf>.

User needs are driving the need for better, bigger and more stable resources, with improved delivery methods:

... it is also necessary for cultural institutions to go beyond the provision of mere databases of disparate objects and intellectual items, to create compelling navigational and learning experiences for end-users and to provide appropriate contexts for use and learning.⁶⁴

The depth of collaboration between LAMS has followed this continuum with *coordination* of small scale experimental projects giving way to *collaboration* to develop federated or aggregated services. The maturity of practice is now starting to require *convergence* to deliver a shared infrastructure for digital curation (Figure 2).

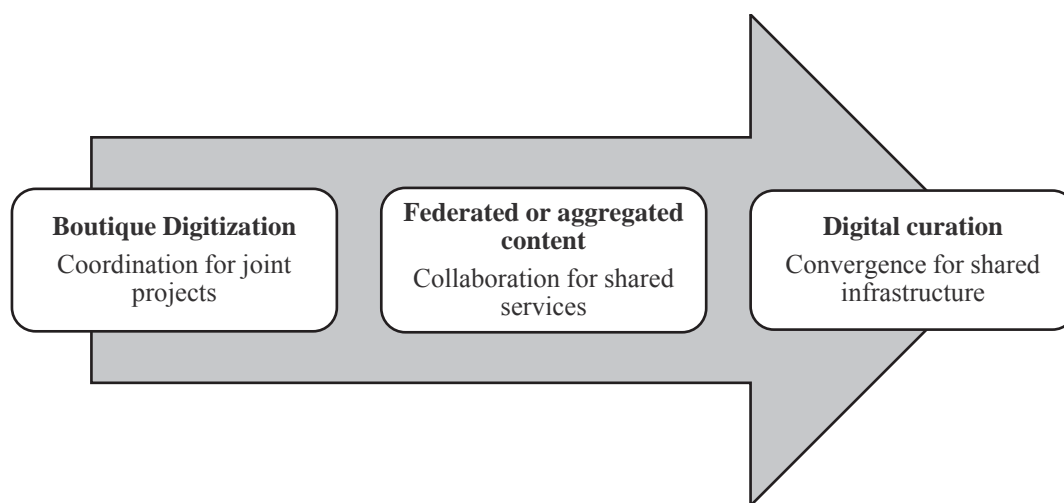


Figure 2. The Digital Creation and Management Continuum

Conclusion

Collaboration across LAMs follows a continuum with increasing levels of trust, integration, commitment and benefits as organizations move through the different stages. Movement between the stages is usually triggered by the need to address a resources gap, and the presence of a change agent to deliver the incentive. Change agents have been prompted by technical advances and development of best practice for digital collection development and management. Not every library, archive or museum is on the same stage in the collaboration continuum, but the imperatives of the digital age mean that most have started the journey, deriving benefits from shared resources and the efficiencies this brings, as trust develops and vision and commitment increase.

⁶⁴ Yeates and Guy, "Collaborative Working for Large Digitisation Projects." citing: M. Nickerson, "Voices: bringing museum exhibits to the world wide web," *First Monday* 7, no. 5 (2002), accessed 31 July 2012, http://firstmonday.org/issues/issue7_5/nickerson/index.html.

Digital Curation

Building an Environment for Success

Jackie R. Esposito

University Archivist and Head, Records Management Services, Penn State University, USA

Abstract

Successful digital curation requires two distinct but dependent sets of operational tools. The first set of tools focus around the archivist/curator identifying and becoming familiar with technical components necessary to implement a viable digital curation operation including the establishment of a digital repository service. The second set of tools, and arguably the more important of the two, is team building. Digital curation at any institution requires developing an ever-changing web of personal, organizational, cooperative, conceptual, and contractual understandings, agreements, and consensus points as well as concurrent relationships. This presentation discusses, in practical terms, the steps, processes, policies, and procedures that need to be established to provide reliable, long-term access to managed digital resources to its designated community, now and into the future. In addition to reflecting on the standard stages of digital curation, it identifies team-building skill sets that will move a digital curation agenda toward both long- and short-term success.

Author

Esposito has been serving as both faculty and administrator for the Penn State University Archives since 2001. She joined the faculty of the University Libraries in July 1991 as Assistant University Archivist for Records Management/Senior Assistant Librarian. She was tenured and promoted to Associate Librarian in 1998. Esposito co-authored Penn State's *ElectRar: Electronic Records Repository Specifications Report*, presented several conference papers on team management of born-digital records utilizing the Matryoshka (Russian Nesting) dolls theory and served on the development team for the Society of American Archivist's Digital Archivist Certificate Curriculum Program.

1. Introduction

Successful digital curation requires two distinct but dependent sets of operational tools. The first set of tools is focused on the archivist identifying and becoming familiar with technical components necessary to implement a viable digital curation operation including the establishment of a digital repository service. The second set of tools, and arguably the more important of the two, is team building.

Digital curation at any institution requires developing an ever-changing web of personal, organizational, cooperative, conceptual, and contractual understandings, agreements, and consensus points as well as concurrent relationships. This paper discusses, in practical terms, the steps, processes, policies, and procedures that need to be established to provide reliable, long-term access to managed digital resources to its designated community, now and into the future. In addition to reflecting on the standard stages of digital curation, this presentation will identify team-building skill sets that will move a digital curation agenda toward both long- and short-term success.

2. Methodology and Scope

This paper reflects on observations from the Society of American Archivist's Digital Archivist Specialization one-day workshop of the same name, taught by this author, dedicated to identifying both the minimum digital curation skill set in conjunction with necessary team building philosophies to begin a program for institutional born-digital archives, records management, and long-term digital preservation. The workshop, as currently designed, establishes five stages of digital curation: 1) **Acknowledge**—the local scope of the issue and its long-term responsibilities; 2) **Act**—initiate “S.M.A.R.T.”¹ projects; 3) **Consolidate**—segue from projects (temporary) to programs (permanent) in staffing and funding; 4) **Institutionalize**—incorporating the larger environment and rationalize programs as part of the operational imperative and permanent budgetary structure; and 5) **Externalize**—embrace inter-institutional collaboration and cooperation.

The workshop also covers and discusses opportunities for success, strategic goals for digital curation, establishing working groups and hot teams, basics of managing electronic content, creating a curation management prototype structure, digital curation management essentials, building effective teams and engaged stakeholders, essentials for digital repositories, and measuring success while re-evaluating goals.

The learning outcomes cover five key areas including:

- An understanding of the components of team building and digital curation necessary to begin working towards a curation prototype in your institution;
- Identification of areas to invest in locally to build knowledge and skills to meets the needs of a digital repository program at your institution;
- Reviewing existing digital repository characteristics that best illustrate roads to success;
- Access to resources, guides, models, and best practices relevant to the digital curation/repository landscape; and
- Identifying and establishing relationships within your organization to achieve a digital archives repository program.

This paper will evaluate the workshop's goals and learning objectives in light of their applications in a real-world environment by utilizing follow-up survey technology for workshop participants. In addition, the paper will reflect on specific team building relationships that consistently work in institutional environments, specifically working with information technology professionals or coordinating scholarly publications with archival repository management. The guidelines developed for creating an environment for success will be quantitatively evaluated as well as qualitatively summarized utilizing narrative focus group summaries.

3. Background on Digital Archives Specialist (DAS) Certificate Program

The Society of American Archivists undertook the development of a certificate program² in 2010 by charging a Digital Archives Continuing Education Task Force to develop a detailed curriculum on digital archives. The Task Force was chaired by Geof Huth (New York State Archives) and included Mahnaz Ghaznavi (Loyola Marymount University), David Kay (Little Airplane Productions), Helen Tibbo

¹ S-M-A-R-T is a mnemonic that stands for Sensible, Manageable, Achievable, Reasonable and Trackable. The point of SMART planning is creating an environment where projects are do-able within a specific period of time.

² www2.archivists.org/prof-education/das

(University of North Carolina/Chapel Hill and this author. The Task Force reviewed existing course offerings and identified curriculum components for awarding a certificate. The proposed curriculum was distributed to sixteen (16) SAA sections and roundtables for comment in March 2011. The Curriculum was edited and revised before being adopted by the SAA Council at its May 2011 meeting.

The DAS curriculum is structured utilizing four tiers of study:

Foundational Courses: essential skills needed to manage digital archives (geared toward practitioners);

Tactical and Strategic Courses: skills needed to make significant changes to develop a digital archive or work managing electronic records (geared toward managers);

Tools and Services Courses: skills needed to work with specific tools and services utilized in managing digital archives (geared toward immediate use);

Transformative Courses: skills needed to change and transform institutions into full-fledged digital archives (geared toward administrators).

A participant has to successfully complete nine (9) required courses from the four tiers and passed the comprehensive exam within a twenty-four (24) month period to receive the DAS certificate. Individual course examinations, multiple choice questions, are provided online immediately after the course is offered.

The curriculum, which features eight (8) foundational courses, eleven (11) tactical/strategic courses, four (4) tools and services courses, and three (3) transformational courses, is based on seven (7) core competencies including:

- Communicating requirements, roles and responsibilities related to digital archives
- Understanding the nature of records in electronic form
- Formulating strategies and tactics for appraising, describing, managing, organizing, and preserving digital archives
- Integrating technologies and tools
- Planning for the integration of emerging technologies
- Curating, storing, and retrieving original masters and access copies
- Providing dependable organization and service

The DAS certificate comprehensive examination, under development at this writing, will be offered at various locations and will cover the spectrum of the curriculum offerings. The DAS certificate is valid for five (5) years. DAS certificate holders may renew their certificates by successfully completing courses or by testing out. SAA is scheduling courses in collaboration with regional organizations and associations as well as on-demand by individual host institutions. As of August 31, 2012, over eight hundred archivists have taken at least one DAS course and several have completed all the coursework necessary to qualify to take the comprehensive exam.

4. Digital Curation: Creating an Environment for Success

4.1 Course Development and Goals

Following the adoption of the DAS certificate program in May 2011, this author began developing the curriculum and learning outcomes for a one-day workshop which would combine two skills sets: 1) digital curation essentials and 2) team building. The workshop is geared to the archival professional

whose institution is **beginning** to discuss digital content management for business records, publications, archival content, research data, and other uses.

The overall goals of the workshop are to encourage participants to “stop waiting and start” pro-active engagement at their local institution. Archivists need to stake a claim in the production cycle, preferably at the beginning of the records life cycle, during its management and use, and through its transfer and preservation. Without a significant infusion of new funds, archival institutions will have to retrain and repurpose existing staff, both internal to their operation and throughout their institution. Archivists should, therefore, focus on being doers within the digital repository process, whenever possible. At all times, archivists should consider digital curation collaborations, internally and externally, as positive developments toward future caretaker responsibilities. By following these recommendations, archivists can actualize collaborative engagements to their benefit for the long-term preservation of the digital object.

4.2 Course Outline

The course is broke down into eleven (11) segments. Each segment is created with its own internal learning outcome and exercises. The segments are:

- ***Opportunities for Success:*** Advocacy – Who are your partners? How do you approach them? What is the optimum working climate? How do you establish an e-records program?
- ***Identifying Strategic Goals for Digital Curation:*** What can be accomplished in six months? One year? Three years? Five years? Developing tactical, operational plans to reach set goals.
- ***Establishing Working Groups and Hot Teams:*** Identify participants and elements of productive teams
- ***Over-arching Focus of Digital Curation:*** What is the Message? How do you deliver it? Prioritizing digital curation for the institutional IT community for the long-term.
- ***Basics of Managing Electronic Content:*** Identifying operational, legal and regulatory, and technological challenges.
- ***Creating a Curation Management Prototype Structure:*** Utilizing the Open Archival Information System (OAIS) Reference Model and the Digital Curation Lifecycle Model to develop a repository structure.
- ***Digital Curation Management Essentials:*** Guidelines, Standards including all ISO requirements, current operational tools and software products, networking and connectivity, and relationship agreements.
- ***Building Effective Teams and Engaged Stakeholders:*** Identifying specific stakeholders, roles and responsibilities, and sustaining inter-departmental operations.
- ***Essential for Digital Repositories:*** Creating a mission statement, policy documents, service agreements, and technical infrastructure.
- ***Bringing the Message to the Masses:*** Providing education, creating a community of users, recruiting advocates, and producing publications to promote the program.
- ***Measuring Success and Re-Evaluating Goals:*** Assessment models and surveys.

4.3 Learning Outcomes

The workshop operates under five (5) specific learning outcomes. The outcomes are:

- An understanding of the components of team building and digital curation necessary to begin working towards a curation prototype in your institution.
- Identification of areas to invest in locally to build knowledge and skills to meet the needs of a digital repository program.
- Review existing digital repository characteristics that best illustrate roads to success.
- Access to resources, guides, models, and best practices relevant to the digital curation and repository landscape.
- Identifying and establishing relationships within your organization to achieve a digital archives repository program.

The workshop focuses on reiterating the core principles of three (3) basic archival theories and practices as delineated in the SAA Code of Ethics, specifically, a) authenticity and integrity—archivists strive to preserve and protect the authenticity of records in the holdings by documenting their creation and use in hard copy and electronic formats; b) access—archivists strive to promote open and equitable access to their collections; and c) security—archivists protect all documentary materials for which they are responsible. Although each category of the Code of Ethics could be highlighted as a part of the workshop focus, time restraints prevent that much compression. It could, however, be part of the marketing strategy for selling the importance of a digital curation program.

In addition, the workshop participants are required to learn and understand the core requirements for digital archives which delineate eight basic structural guidelines for success. These requirements include:

- Ensuring integrity, authenticity and usability of digital objects
- Operating within an efficient and effective policy framework
- Acquiring and maintaining contractual, legal, historical, fiscal, and organizational rights
- Committing to maintenance of digital objects
- Maintaining the repository permanently *with sufficient resources* (author's emphasis)
- Acquiring and ingesting digital objects based on stated criteria and policies
- Creating and maintaining requisite metadata about digital objects during preservation, production, access and usage processes
- Fulfilling dissemination requirements

4.4 Student Expectations

Prior to offering each workshop, registered attendees are asked to identify their learning objectives. These summaries are shared with the instructor in anticipation of allowing the instructor to address them with the workshop participants as well as help define the scope of the instruction². The expectations for *Digital Curation: Creating an Environment for Success* has, to date, fallen into four learning categories:

1. Practical strategies for engaging stakeholders
2. Identify resources for best practices including terminology standards
3. Methods for providing access to born-digital content
4. Better understanding of ways to establish a repository program

Interestingly, both in the course of offering the workshops and in the assessments, it has become clear that scale is an issue. Solutions that work for large universities may not be applicable for smaller local historical societies. One of the foci for software developers and vendors should be to address these scale concerns for all phases of the born-digital lifecycle from ingest through management to access and preservation. What is the best way to offer services, tools, and applications for both state and local governments when staffing, budgets, and institutional support differ significantly? Scale can be an issue during instruction as well. The resources, especially staff, available to address digital curation issues definitely affect opportunities for success.

Several workshop attendees expected the workshop to be “extremely useful to help in strategic planning for preservation of my center’s digital resources.” These expectations expose a basic issue within the entire DAS program, that is, the professional confusion between concepts such as digital curation, digital preservation, electronic records management and digital repository development. This confusion requires that the instructor spend a portion of every workshop clarifying terms, concepts, and delineating the scope of the particular workshop across the spectrum of digital archival activities. While this confusion may be attributable to the current state of archival training and expectations, this author believes it is a significant issue across institutions as well since the concepts and their definitions change across professions and professional associations. Clarifying the language of the project and/or program at a specific institution is significant communication hurdle for practitioners as well as workshop instructors. Stabilizing vocabulary is a first, and critical, step in team building success. Communication fails if language variations are not address in advance.

Another major stepping stone is revealed in expectations from participants that include specific learning goals such as “advantages/disadvantages of using scanners vs. digital cameras for image capture” or “evaluations of digital asset management systems.” While this particular workshop does discuss and identify specific software and their vendors, at no time does the workshop go into the type of operational details that these participants seem to desire. Rather than understanding the language to have a vendor conversation, these participants clearly want specific guidance for day-to-day operational specifications. There are DAS courses that offer this kind of detail, this workshop does not.

4.5 Specific Group Activities

The workshop opens with an ice breaker activity which requires the participants to find attendees who fit specific categories. Titled ***Have you ever?*** the questionnaire asks about living overseas, singing karaoke, eaten frog legs, seen a polar bear, etc. There are twenty categories and several of the activities are written in such a way as to require clear communication skills. For example, ***have you ever flown a plane?*** The question does not ask “have you ever flown in a plane?” The difference between the two questions is extensive. Every single workshop taught to date has misread that question which opens up a huge avenue for discussion about language meaning and understanding.

The activity lasts 7-10 minutes. During the review period, two lessons are discussed: 1) if word constructs, such as the one above concerning the airplane, are so important, how would that skill translate in team building for digital curation? 2) How many attendees, colleagues, participants share traits, activities, interests that are new to you? Sharing these newly discovered traits expands the communication paths, both professionally and personally.

A second activity involves breaking entire group into smaller cohorts of four-to-six participants. The smaller groups are given the opportunity to identify seven items necessary to survive on an island.

The type of island, the length of time on the island, and available resources are not specified. Participants have to prioritize and strategize to survive. Certain tools and skill sets are necessary while others are “luxuries” when survival is the goal. This activity is translated from an island environment to establishing priorities for digital survival. It requires the participants to create strategic plans for their digital curation development, maintenance and permanence. Interestingly, certain characteristics always rise to the top of the survival list: staffing, equipment, expertise and funding. These characteristics are not surprising or unexpected. They are, however, brought to the forefront of the participant to allow them to begin identifying strategies for acquiring necessary tools.

Other activities utilized during the workshop to engage participants are geared toward developing a working outline for use when attendees return to their individual workplaces. For example, depending on time, participants will develop a marketing, promotional campaign. This author’s favorite to date compared digital curation to fruit preservation via canning. The visual comparisons and the use of food as a motivator are strikingly affective. There are no individual activities or tasks. Since the workshop is focused on team building, group activities are required.

4.6 Student Evaluations

After the workshop has been completed, participants are asked to assess the course “from the standpoint of what you gained from the experience.” The participants are also asked to rate the “methods and materials relative to their value in accomplishing the course.” In addition, they are asked to rate the instructor, identify the most valuable methods and materials, identify aspects of the course that should be changed, and provide any additional comments.

4.6.1 Question 1

In the question relative to assessing the course for experience gained, there are eight segments:

- Components of team building
- Pinpoint areas for local investment
- Review existing repository characteristics
- Access to resources and best practices
- Recognize and establish organizational relationships
- New knowledge/skills acquired
- Likelihood of applying concepts learned
- Expectations met

To date, the participants have valued the balance in the workshop between technical born-digital materials management and team building as they relate to real-world experiences. Generally, the highest scoring questions have been *components of team building* with a 4.68 out of 5 and *pinpointing areas to invest in locally to build knowledge and skills* with an average of 4.47 out of 5.

The lowest scoring question in this section has been *new skills acquired* with an average 4.09 out of 5 and *likelihood of applying concepts to your work* with an average 4.19 out of 5. These scores are alarming in that archival professionals are taking (and paying for) professional development that they readily report may not be able to apply in their daily jobs. The survey does allow participants to discuss specific operational hurdles. These hurdles are revealed during group discussions but they usually require one-on-one solutions local to the institution.

4.6.2 Question 2

The question asking participants to rate methods and materials specifically asks for feedback on clarity of handouts, content of handouts, exercise/group discussions, clarity of visual aids and content of visual aids. The highest scoring questions in this group has been *exercises/group discussion* with a 4.45 out of 5 score and *clarity of visual aids* at 4.41 out of 5. These scores are clearly concrete recommendations for instructors relative to the needs of the workshop participants to interact with one another and visually see examples of content as it is being discussed. Workshops must be vibrant and relatable to real-world scenarios to work with digital curation practitioners.

The lowest scoring question in this group has been *clarity of participant handouts* with a 4.27 out of 5 and *content of participant handouts* at 4.31 out of 5. These responses are a tad confusing when compared with the open ended comments in questions four through six which often include complaints about the size and density of PowerPoint slides. Workshop participants either do not value the handouts during the workshop due to their preference for group discussions or expect handouts to serve as a post-workshop tool in which case the clarity and content is evaluated in its reflection at a later date. Additional study is required to understand the respondent's answers. It might be possible for SAA to restructure the question to provide the workshop instructor with more detailed information and guidance.

4.6.3 Questions 3

The third set of questions in the assessment asks for participants to rate the instructor on knowledge of topic, preparation, ability to handle questions, and presentation skills. Fortunately for the workshop instructor/author, these scores have been consistent at 4.91 out of 5. One of the most positive aspects of both these questions and comments in questions four through six has been the participant's emphasis on appreciation of the instructor's positive attitude, especially "in light of the enormity of the topic."

4.6.4 SUMMARY OF ASSESSMENT RESPONSES

	Session 1	Session 2	Session 3	Session 4	Session 5	Totals
Question One: Assess the course from the standpoint of what you gained from the experience						
Team Building and Digital Curation Components	4.65	4.66	4.81	4.55	4.71	4.68
Local Investments	4.39	4.36	4.56	4.64	4.38	4.47
Existing digital curation characteristics	4.03	4.45	4.63	4.18	4.48	4.35
Access to resources, guides, best practices	4.26	4.59	4.63	4.36	4.67	4.5
Recognize and establish relationships	4.61	4.69	4.69	4.45	4.57	4.6
New knowledge skills acquired	3.74	4.34	4.06	3.91	4.29	4.09
Likelihood of applying concepts	4.45	4.17	4.06	3.82	4.43	4.19
Expectations met	4.26	4.55	4.44	4.27	4.48	4.4
Overall Score	4.3	4.48	4.48	4.27	4.5	4.41

Question Two: Rate the methods and materials relative to their value in accomplishing the course.

Clarity of participant handouts	4.2	3.96	4.33	4.45	4.43	4.27
Content of participant handouts	4.13	4.25	4.4	4.27	4.48	4.31
Exercises/Group Discussions	4	4.38	4.63	4.36	4.86	4.45
Clarity of Audio-visual aids	3.96	4.38	4.5	4.78	4.43	4.41
Content of Audio-visual aids	4	4.57	4.57	4.33	4.38	4.37
Overall Score	4.06	4.31	4.49	4.44	4.51	4.36

Question Three: Rate Individual Instructor

Knowledge of topic	4.81	4.97	5	4.91	4.95	4.93
Preparation	4.77	4.97	4.94	5	4.81	4.9
Ability to Handle Questions	4.77	4.97	4.94	5	4.95	4.93
Presentation Skills	4.74	4.9	4.94	4.91	5	4.9
Overall Score	4.77	4.95	4.95	4.95	4.93	4.91

4.6.5 Questions 4-6

The last three questions on the assessment are open-end, participants can comment on any aspects of the workshop in a narrative format which in many ways is the most valuable feedback for the instructor. One of the participants summed it very nicely, “stepping back to see the big picture mission and strategy for digital curation and encouraging us here to build teams with our archival colleagues.” The most consistent positive comments fall into specific areas:

- Small group exercises and discussions are widely valued
- Participants appreciate practical tools, real-life examples and best practices
- Focus on developing realistic strategic and action plans is appreciated
- Exercises to develop critical thinking on the topic should be mandatory
- Policy outlines, discussions and strategies for repository management
- Gaining a general overview of digital curation and repository processes is critical

Eric M. Kimura attended to Ventura, CA workshop and summarized his experience, “very comprehensive picture of digital curation management.” Another attendee summarized the workshop goals quite well, “knowing the steps for getting started with digital curation was invaluable.” While archival and information technology professionals will admit that the digital curation conversation has been going on for quite a few years, the reality is that most institutions are addressing the problem in fits and starts. Foundational workshops that identify the key issues and address the communication needs across departmental boundaries are critical at this juncture.

Among the segments of the workshop that participants would change are those areas which, unfortunately, cover the most technical areas of the born-digital life cycle management such as understanding the OAIS model and the digital curation lifecycle. The other prominent emphasis for change extends to the need to collaborate in areas that force participants to extend themselves outside their comfort zone to interact with professionals from other fields such as information technology. Conclusion: team building is hard; digital curation is complicated; success requires standing up to both challenges, facing them head on, and moving forward at each juncture.

5. Lessons Learned

Erica Bondreau, JFK Library Digital Archivist and DAS student blogger, commented after attending the Boston workshop, “this course represented a return to the concerns surrounding born-digital material.” The overriding metaphor for the workshop is the oft-utilized “you can’t eat an elephant whole.” Large, complex programs, such as digital curation, need to be broken down into manageable pieces to achieve successful outcomes. Bondreau extended that metaphor in her blog, “an elephant, once broken apart, is best eaten with friends.” She is correct, digital curation “cannot be managed by one archivist alone. Partnering is key to your success.”

The major lesson learned in conducting these workshops has been the overwhelming value of real-world processes, best practices and examples. Each workshop allows group discussions around topical concerns. In every case, one or two major projects currently being undertaken became the focal point for practice. Whether it be the large data set curation efforts at Purdue, digital programming at WGBH in Boston, state government records in Sacramento, local property records in Lancaster, or genealogical family records in Provo, the issue revolves around a central core—how can I make this work at my institution? Often without new resources, staff or funding, these professionals are hungry for real world, affordable solutions. The workshop goals are designed to address those issues specifically.

The other lessons have been technical: are eight hours too long? Too short? Are PowerPoint’s too text-rich? Not enough? Can small repositories learn from larger ones and vice versa? These issues can be addressed on-site during a specific workshop and with a specific set of participants. The value of the pre-attendance questionnaires is immeasurable when considering these issues.

6. Planning for the Future

There are two adjustments made to every workshop in advance of its offering: 1) Update available tools and 2) Reevaluate best practice examples. It is incumbent on the workshop instructor to stay current. The digital curation landscape changes every day. Since its core is technology, the options are constantly in flux. It is imperative upon the instructor to adjust the take-aways from the workshop so that they are as up-to-date as possible. It is also critical for the instructor to always focus on core competencies. Tyler Walter delineated these in his 2011 Association of Research Libraries report entitled *New Roles for New Times*. He stated “collaborative or community-based approaches to digital curation are likely to be more effective and sustainable...when following these recommendations:

- Stop waiting and start proactive engagement locally
- Stake a claim in the production cycle
- Start retraining and repurposing staff
- Be a doer, not a broker, wherever possible
- Consider digital curation collaborations
- Actualize collaborative engagement

Archival professionals understand the necessity to host digital content and curate the core components of their work. What they require from the Society of American Archivists and other professional associations is educational guidance in the form of workshops, tutorials, publications, and solid advice. Digital curation is too complex and expensive to experiment and fail miserably. Therefore, many professionals expect assistance that will help them succeed and subsequently save our digital heritage.

Notes

1. Comments noted in this section have been extracted from SAA workshop evaluation summaries. Workshops, during 2012 have been held on the following dates in the detailed archival regions:

- Boston, MA, January (31 respondents)
- Ventura, CA, April (29 respondents)
- Provo, UT, June (16 respondents)
- University Park, PA, June (11 respondents)
- West Lafayette, IN, July (21 respondents)

Total number of surveys reviewed: 108

Additional workshops have been scheduled for New Orleans, LA (October), Austin, TX (December), and Philadelphia, PA (January 2013). It will be interesting to see if the response trends shift.

References

- AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*, 2012.
<http://www2.lib.virginia.edu/aims/whitepaper>.
- Bondreau, Erica. "Diary of a DAS Student." February 8, 2012. <http://dasstudentdiary.blogspot.com/>.
- Carroll, Laura, Erika Farr, Peter Hornsby, and Ben Ranker. "A Comprehensive Approach to Born-Digital Archives." *Archivaria* 72 (Fall 2011): 61-92.
- Cunningham, Adrian. "Good Digital records Don't Just 'Happen': Embedding Recordkeeping as an Organic Component of Business Processes and Systems." *Archivaria* 71 (Spring 2011): 21-34.
- Curation Lifecycle Model*. Digital Curation Centre, 2008.
- Digital Preservation Coalition. *Preserving Email*. www.dpconline.org/newsroom/latest-news/805-email-tomorrow.
- Digital Standards at the Library of Congress*. www.loc.gov/standards/.
- Electronic Records Management Guidelines*. Minnesota Historical Society.
http://www.mnhs.org/preserve/records/electronic_records/erguidelinetoc.html.
- Erway, Ricky. "Defining 'Born Digital'." OCLC Research, 2010.
www.oclc.org/research/activities/hiddencollections/borndigital.pdf.
- Erway, Ricky. "You've got to walk before you can run: First Steps for Managing Born-Digital Content received on Physical Media." OCLC Research, 2012.
<http://www.oclc.org/research/publications/library/2012/2012-06.pdf>.
- Harvey, Ross. *Digital Curation: A How-to-do-it Manual*. Facet Publishing, 2010.
- Hwse, Patricia, Michael J. Giarlo, Michelle Belden, Kevin Clair, Daniel Coughlin, and Linda Klimczyk. "Building a Community of Curatorial Practice at Penn State: A Case Study." *Journal of Digital Information* 13, no.1 (2012). <http://journals.tdl.org/jodi/article/viewArticle/5874/5880>.
- ISO Standard for Trustworthy Digital Repositories*. ISO16363:2012.
- Nelson, Naomi et al. *Managing Born-Digital Special Collections and Archival Materials*, Spec Kit #329. Association of Research Libraries, 2012.

- O'Meara, E., and Tuomala, M. "Finding Balance between Archival principles and Real-Life Practices in an Institutional Repository." *Archivaria* 73 (Spring 2012): 81-103.
- Phillips, Meg. "More Product, Less Process for Born-Digital Collections: Reflections on CurateCamp Processing." August 22, 2012, <http://blogs.loc.gov/digitalpreservation/2012/08/more-product-less-process-for-born-digital/>; "Preserving Access to our Digital Future: Building an International Digital Curation Curriculum." <http://www.ils.unc.edu/digccurr> and <http://www.ils.unc.edu/digccurr/products.html>.
- Prom, Christopher. "Making Digital Curation a Systematic Institutional Function." *The International Journal of Digital Curation* 1, no. 6 (2011). <http://www.ijdc.net/index.php/ijdc/article/view/169>.
- Prom, Christopher. *The Practical E-Records Method*. <http://e-records.chrisprom.com/recommendations/>.
- "Special Issue on Special Collections and Archives in the Digital Age," *Research Library Issues: A Quarterly Report*, RLI 279, June 2012.
- Sustainability of Digital Formats: Planning for Library of Congress Collections*. www.digitalpreservation.gov/formats/.
- Toolkit for Managing E-Records*. National Archives and Records Administration. <http://toolkit.archives.gov>.
- Walters, Tyler. *New Roles for New Times*, 2011. <http://www.arl.org/rtl/plan/nrnt/nrnt-dcwebcast.shtml>.

Tools

- Archive-It Web Archiving Service: <http://www.archive-it.org/>
- Archivematica: <http://archivematica.org>
- BitCurator Tool: <http://www.bitcurator.net>
- California Digital Library's Web Archiving Service: <http://webarchives.cdlib.org>
- Curator's Workbench (ingest and description tool): <http://www.lib.unc.edu/blogs/cdr/index.php/2010/12/01/announcing-the-curators-workbench/>
- DROID (Digital Record Object Identification): <http://sourceforge.net/projects/droid/>
- Duke Data Accessioner (Ingest tool): <http://library.duke.edu/uarchives/about/tools/data-accessioner.html>
- Forensic Toolkit Imager: <http://accessdata.com/support/downloads>
- HTTrack (local website copying app): <http://www.httrack.com>

Célébrations institutionnelles

Événement catalyseur propice à l'implantation d'un projet de conservation du patrimoine numérique permettant de réunir les acteurs d'intérêts divergents

Patricia Forget

Archiviste universitaire et muséologue, Canada

Résumé

Comment profiter d'une occasion et d'un budget de célébrations, projet rassembleur, pour implanter un projet de conservation du patrimoine numérique, et ce, malgré les intérêts divergents entre les créateurs et les conservateurs? J'expliquerai les avantages et défis de participer à de tels événements qui permettent aux responsables de la conservation de mettre de l'avant les projets d'acquisition et de traitement du patrimoine numérique en trouvant et en implantant une solution technologique qui puisse répondre à une multitude de challenges, dans le but ultime de mettre en valeur ce patrimoine et de le conserver. Le texte prend position pour la nécessité du regard humain sur la validation de tout patrimoine numérique.

Auteur

Archiviste universitaire à l'Université du Québec en Outaouais (UQO) depuis 2008, Patricia Forget est chargée de cours en cybermuséologie à l'École multidisciplinaire de l'image de l'UQO depuis 2010. Elle est membre du sous-comité des archivistes du Conseil des recteurs et des principaux des universités du Québec (CREPUQ) depuis 2009, ainsi que du Comité des usagers responsables de la gestion documentaire (CURGD) de l'Université du Québec depuis 2010. Elle siège en tant que conseillère au Comité Région-Ouest de l'Association des archivistes du Québec (AAQ) depuis 2008. Elle a aussi participé en tant que chef de l'équipe des archives qui a mené divers projets de numérisation, dont celui du Musée canadien des civilisations (MCC) en 2007 et 2008, un projet financé en partie par Culture canadienne en ligne (CCE), une initiative du ministère du Patrimoine canadien.

1. Introduction

Contrairement à la croyance populaire, les anciennes méthodes de documentation et de classement conçues et appliquées par des générations de scientifiques depuis le Siècle des Lumières ne sont pas aussi désuètes qu'on aimerait bien le prétendre aujourd'hui. Ces outils de repérage et de conservation des données d'observation pouvaient être utilisés de génération en génération sans que ce savoir-faire ne se perde :

[...] ces sociétés ont globalement donné l'impression aux différentes générations qui se succédaient qu'un progrès d'ensemble était à l'œuvre. Cette idéologie du progrès s'est développée au XVIII^e siècle dans tous les domaines : social, philosophique, politique, littéraire, économique. Au XX^e siècle, elle a été sérieusement mise à mal, mais son déclin a été en quelque sorte recouvert par un autre phénomène : la consommation de masse.¹

L'observation des phénomènes naturels caractérisait une des principales étapes de l'ère de la rationalité. En fait, l'observation par des dessins ou des photographies, la description détaillée de phénomènes et de

¹ J.-F. Ballay, « Paradoxes de la transmission et de l'apprentissage dans un monde radicalement incertain, » *Télescope* 16, n° 1 (2010) : 2-3.

processus permettait aux générations futures de reprendre les données brutes qui se voyaient attribuer un numéro unique grâce auquel il était facile de regrouper l'ensemble des constituantes, quelque fût leur format, pour les analyser par la suite sous un nouvel angle, selon un nouveau regard. Est-ce que ces méthodes de documentation sont encore présentes dans nos milieux de travail, dans nos universités? Est-ce qu'elles ne devraient pas l'être?

Avec l'apparition de la culture de masse, la société de consommation a eu comme résultat, dans un premier temps, l'émergence des technologies. Cette technologie permit d'une part la propagation des médias de masse (photographie, cinéma, radio, télévision, etc.) et d'autre part, facilita le traitement des données² (le calcul, via la calculatrice, puis l'ordinateur qui débute en effectuant uniquement la gestion des données) : « Les médias de masse et le traitement de données sont des technologies complémentaires; elles apparaissent ensemble et évoluent côte à côte, rendant ainsi possible la société de masse moderne. »³ Tout le XIX^e siècle et le XX^e siècle, il y a une distinction claire et définie entre les médias de masse et les technologies pour le traitement des données. Ils sont traités dans deux systèmes différents et de deux manières différentes, ce qui permet un traitement très différent de ces deux types d'informations.

Malgré ces changements technologiques, la méthodologie de documentation héritée du Siècle des Lumières poursuivait son chemin et permettait toujours d'identifier ces nouveaux formats pour la conservation de l'information. Le repérage des médias de masse s'effectuait encore selon la méthodologie développée au Siècle des Lumières, soit l'attribution d'un numéro unique dont chacune des constituantes pouvait bien identifier le type de format, le contenu général, l'année de création de ce document, etc. Il permettait le repérage rapide à l'aide d'outils, mais également la gestion efficace de ce savoir afin de toujours mettre de l'avant l'accessibilité de ces ressources. Pour ce qui est du traitement de données, plusieurs types de rapports avaient été créés afin de rassembler dans un même document l'ensemble des données pertinentes. Certains types de documents devinrent essentiels à la compréhension de tout traitement de données. Parlons, par exemple, des journaux comptables et des états financiers, pour le domaine de la comptabilité; ces types de documents rassemblaient souvent l'ensemble des données pertinentes. À la fin du XX^e siècle et au début du XXI^e siècle se produit un changement majeur : « De machine analytique, tout juste bonne à traiter des nombres à grande vitesse, il est devenu [...] un outil de synthèse et de manipulation médiatiques. »⁴ C'est ce que nous appelons maintenant les nouveaux médias. Avec l'arrivée de ces nouvelles technologies, un nouveau mythe est né, voulant qu'elles permettent de répondre à tous les problèmes d'identification, de repérage, d'accessibilité, et ce, sans l'aide de l'humain! L'ère tant attendue de la déresponsabilisation semblait enfin arrivée. C'est alors que de nouvelles

² « En 1887, le Bureau du recensement américain s'employait encore à interpréter des chiffres de 1880. Pour le recensement de 1890, l'agence adopta des machines de tabulations électriques conçues par Herman Hollerith. [...] Au cours de la décennie suivante, les tabulateurs électriques firent partie de l'équipement normal des compagnies d'assurances, des services publics, des bureaux des compagnies de chemin de fer et des services de comptabilité. En 1911, la Tabulating Machine Company d'Hollerith fusionna avec trois autres compagnies pour former la Computing-Tabulating-Recording Company et en 1914, Thomas J. Watson fut choisi pour la diriger. Dix ans plus tard, son chiffre d'affaires avait triplé et Watson la rebaptisa International Business Machines Corporation, ou IBM. »

L. Manovich, « Que sont les nouveaux médias : Le mythe du numérique, » *Le langage des nouveaux médias* (Dijon : Les Presses du réel, 2010), 93.

³ « La capacité de diffuser les mêmes textes, images et sons à des millions de citoyens, inculquant ainsi des croyances idéologiques communes, était aussi essentielle que celle de garder une trace de leur acte de naissance, du registre de leurs emplois successifs, de leur casier judiciaire ou de leur dossier médical. » Manovich, 91.

⁴ Manovich, 96.

fonctions et/ou de nouveaux services furent mandatés de cette nouvelle responsabilité, et ce, sans avoir souvent les connaissances appropriées pour jouer un tel rôle.

Malheureusement pour les admirateurs de science-fiction, la réalité est tout autre. Au lieu de simplifier les choses, l'arrivée des nouveaux médias crée, au contraire, une confusion entre les données essentielles d'une institution et les données superflues. Cette confusion est d'autant plus réelle que la création de médias de masse et de données n'a jamais été aussi facile, ce qui crée un problème de quantité qui augmentent d'autant plus cet imbroglio. Il est donc encore plus important aujourd'hui de bien différencier ce que l'on doit garder de ce que l'on ne doit pas conserver, et ce, dès leur création, avant même que le chaos ne s'installe. De plus, la vitesse à laquelle il peut maintenant s'installer est vertigineuse. Ce mélange des données essentielles d'une institution avec les médias de masse amène un danger réel au niveau du traitement. Quels sont les documents essentiels au bon fonctionnement d'une institution? Qu'est-ce que nous pourrions éliminer rapidement sans que cela ait d'incidence à court, moyen et long-terme?

La surspécialisation des disciplines, qu'elles soient universitaires, administratives, juridiques ou commerciales nous ramène à une pluralité de méthodes de documentation et de classement digne de la tour de Babel. Les nouveaux services spécialisés dans le traitement ou le support de ces informations ne peuvent travailler en vase clos. Ils doivent constamment se référer aux spécialistes de cette information qui la traitent avec la rigueur qu'elles méritent, ce qui ne peut être fait sans l'œil vigilant de l'humain :

Identifier les domaines connexes critiques pour l'organisation. [...] En surchargeant les bases de connaissances et les systèmes d'information, le risque d'inonder les employés d'information et d'entraîner une détérioration plutôt qu'une amélioration de la situation. L'approche à privilégier commence donc par la définition de priorités visant à ce que les actions entreprises contribuent significativement à l'atteinte des objectifs organisationnels tout en minimisant la boulimie informationnelle.⁵

Voilà pourquoi les réels spécialistes de ces médias de masse et de ces données créées à l'intérieur d'une institution sont avant tout les créateurs eux-mêmes.

Est-ce que nos milieux ont toujours la rigueur nécessaire à l'application de l'entière de ces méthodologies? En fait, contrairement à la croyance populaire, je crois certainement qu'il n'a jamais été aussi important et nécessaire de documenter l'entière du travail des individus, et ce, en se référant aux méthodologies scientifiques, méthodes qui permettaient de diviser les données brutes des analyses et surtout de documenter le processus de collecte de ces données, sans lequel les données n'ont plus aucune valeur.

2. Développement

Au-delà du capital informationnel intimement lié aux systèmes d'informations et aux documents, le capital humain est la « matière première » nécessaire pour valider la qualité et l'intégrité des données, des informations et des connaissances [...] Les incontournables du savoir patrimonial sont des personnes possédant une compétence unique, acquise au cours des ans et adaptée aux exigences de leur environnement professionnel.⁶

⁵ L. Rivard et M.-C. Roy, « Un cycle de rétention des connaissances pour combattre l'amnésie organisationnelle, » *Télescope* 16 (2010) : 67-81, pp. 70-71.

⁶ Rivard et Roy, 68.

Cette constatation peut invariablement s'appliquer à chacune des institutions dans lesquelles nous évoluons dans notre monde contemporain. La pensée magique voulant qu'il y ait un jour un moyen, une technologie, un nouveau média qui puisse remplacer la documentation et surtout la validation par une personne compétente en la matière des processus organiques d'une institution s'avère très dangereuse. Je prends position pour la nécessité du regard humain sur la validation de toute information et je souhaite vous le démontrer par mon étude de cas.

Que cette institution soit publique ou privée qu'elle soit un centre d'archives, un musée ou une bibliothèque, institutions communément dénommées « institutions de mémoire », il importe de bien mettre en lumière cette réalité. Et ma présentation se veut une prise de position en ce sens. J'insiste sur cet aparté parce que trop souvent les créateurs de l'information d'une institution soulignent qu'ils n'ont pas à effectuer ce type de travail justement parce qu'ils ne font pas partie d'une institution de mémoire ou d'un service de mémoire, et que leur fonction n'implique pas la validation de la qualité et de l'intégrité des données, des informations et des connaissances. Mais, ont-ils vraiment le choix de ne pas s'impliquer?

3. Reconnaître

En avril 2009, j'ai été invitée à participer aux célébrations du 30^e anniversaire de l'Université du Québec en Outaouais (UQO). Le comité organisateur regroupait du personnel des communications, du corps professoral, des ressources humaines, de l'informatique et de la gestion documentaire, ce qui, au Québec, inclut aussi les archives.⁷ Pour moi, c'était l'occasion toute désignée de faire connaître les besoins que j'avais déjà analysés quelques mois auparavant, à mon arrivée en poste.⁸ En effet, l'analyse de besoins en gestion documentaire que j'avais réalisée avec l'aide du secrétaire général avait soulevé une problématique importante sur le plan de la gestion du patrimoine photographique, qu'il fût analogique ou numérique. Aucune photographie n'avait jamais été transférée au Service des archives de l'Université, à part celles du Service de l'audiovisuel (SAV).⁹ Lors d'une entrevue, les créateurs et les utilisateurs de ces photographies, particulièrement la Direction des communications et du recrutement (DCR), m'avaient interpellée à ce sujet. Ils ne les retrouvaient plus pour leurs besoins de diffusion. Quand, enfin, ils les retrouvaient, celles-ci étaient souvent décontextualisées et sans identification (nom de l'auteur, événement, date, etc.).

À titre de membre du sous-comité mémoire, comité créé pour réaliser les projets reliés à la mise en valeur du patrimoine de l'Université, et ce, en collaboration avec l'École multidisciplinaire de l'Image (EMI), j'ai suggéré que le 30^e anniversaire pourrait devenir une occasion hors pair pour analyser l'ensemble du patrimoine photographique de l'Université de façon globale, puisqu'il n'avait jamais été traité dans son ensemble. En effet, les célébrations institutionnelles importantes deviennent souvent l'occasion unique pour une institution de réaliser des projets importants qui permettent de réaliser des pas de géants dans l'établissement en établissant de nouvelles collaborations et en solidifiant celles du passé.

⁷ C'est une particularité québécoise. En effet, la particularité de l'archivistique québécoise est qu'elle regroupe sous une même profession le responsable de la gestion documentaire, ou « record management », et l'archiviste.

⁸ Deux sous-comités furent créés afin de permettre une meilleure exécution des projets : le sous-comité mémoire et le sous-comité livre.

⁹ SAV n'existe plus aujourd'hui à l'UQO. Il a été fusionné en 1989 au Service informatique appelé maintenant Service des technologies de l'information.

Le rayonnement de tels événements permet également de sortir de l'ombre des services essentiels de l'Université tels que les services de mémoire, en l'occurrence le Service des archives.

L'importance d'une sélection appropriée est vraiment l'élément clé dans tout projet de numérisation et de conservation. Lorsque l'on embarque dans un projet de célébrations institutionnelles, une grande part d'inconnu s'installe, surtout lorsque le corpus n'a jamais été analysé dans son ensemble et qu'aucun traitement approprié n'a été effectué—ce qui était mon cas. En effet, comment arriver à sélectionner du matériel qui n'a jamais été traité ni documenté? Deux cas de figure s'observent : le fait qu'une institution soit petite et jeune permet d'accéder plus aisément et rapidement à l'ensemble des responsables des unités administratives, d'en faire le tour, d'identifier le patrimoine vivant (personnel de plus de 25 ans de service) et du coup d'arriver en peu de temps à bien circonscrire le corpus qui nous permettra d'arriver à des résultats concluants.

Dans des institutions de mémoire telles que les musées nationaux, ce type de projet pourrait prendre plusieurs années de recherche et demander la collaboration de nombreux intervenants spécialisés, mais permettrait de réaliser un projet de qualité à partir d'un corpus souvent traité et déjà mis à la disposition des chercheurs.¹⁰ En revanche, dans le cas d'une institution qui ne se considère pas comme une institution de mémoire et surtout qui, de par sa nature, se retrouve plus dans un mécanisme de survie que dans une dynamique de création et surtout de documentation de leur patrimoine, l'enjeu est bien plus complexe. En fait, comment arriver à faire naître rapidement ce sentiment de longévité dans une organisation encore jeune, mais qui atteint, par le fait même une certaine maturité?

Une piste de solution est que le traitement de ce corpus devienne la base nécessaire à la réalisation de projets de qualité pour cette célébration, mais également pour bien établir les bases de l'accès du patrimoine passé en créant un pont pour l'avenir. Bien circonscrire un corpus donné peut s'avérer un travail de longue haleine, que ce soit dans le monde de l'analogique ou dans le monde du numérique. Pour certaines personnes qui arrivent à la fin de leur mandat, ce passé est encore bien présent et il permet souvent de faciliter la route, trouver les raccourcis qui permettent d'arriver à trouver le corpus essentiel, documenté dans un délai raisonnable. Circonscrire le corpus, voilà la première étape.

C'est ainsi que deux corpus se sont avérés incontournables : celui de DCR et celui du SAV. Pour ce présent texte, nous nous concentrerons sur celui des communications. D'abord nommé Service d'information et des relations publiques (SIRP), le Service des communications créé en 1975 est devenu en 2003 la DCR. Alors, comment profiter d'une occasion et d'un budget de célébrations, projet rassembleur, pour implanter un projet de conservation du patrimoine numérique, et ce, malgré les intérêts divergents entre les créateurs et les conservateurs?

Peut-être justement en trouvant ce qui nous rallie. C'est souvent en y allant avec les besoins déjà exprimés par les principaux utilisateurs que nous arrivons à mettre en branle une bonne relation. Il est clair que pour les créateurs et les conservateurs, l'idée de retrouver rapidement une photographie est gagnante. C'est donc à partir de la dénomination (identification appropriée) que nous avons amorcé notre

¹⁰ Ayant soutenu les conservateurs et auteurs dans leur recherche de documentation pour l'exposition et le livre du 150^e anniversaire, j'avais été directement impliquée à chacune des étapes de conception et de réalisation de l'exposition. Voir : Musée canadien des civilisations (MCC) (2006, avril). *Exposition : 150 ans de culture, de collections et de découvertes au Musée canadien des civilisations*. Version revue et corrigée (2007, avril), consultée le 14 août 2012, http://www.civilisations.ca/cmce/exhibitions/cmce/150/m150_01f.shtml; Vodden, Christy, et Ian Dyck. *A World Inside: A 150-Year History of the Canadian Museum of Civilization* (Gatineau, Québec : Canadian Museum of Civilization Corporation, 2006), 104 p.

relation. Cette relation avait été établie dès le mois de mai 2009, à la suite de mon analyse de besoins.¹¹ Rencontres hebdomadaires de service où j'ai été invitée à venir effectuer une présentation à l'ensemble des employés du service afin de venir leur parler de préservation et leur proposer un modèle de numérotation que je voulais compatible avec leurs besoins et leur façon de faire. De cette réunion a découlé une procédure d'identification, dont voici des extraits :

Niveau du dossier : **Collationdesgrades_Ph0_01092009_PF_0001**

1. Code ou nom de l'événement : **Collation des grades**
2. Type de production : **Pro, Pub, Pho** (Projet ou Publicité ou Photographie)
3. Jour/mois/Année : **01092009**
4. Code d'identification du créateur : **PF** (Patricia Forget)
5. Numéro de dossier : **0001**

Important : Ce dossier doit inclure l'ensemble des fichiers : images, vidéos, texte, etc.

Niveau de chacune des pièces du dossier :

Collationdesgrades_Ph0_01092009_0001_PF_Recteur_0001

1. Code ou nom de l'événement : **Collation des grades, ...**
2. Type de production : **Pro, Pub, Pho** (Projet ou Publicité ou Photographie)
3. Jour/mois/Année : **01092009**
4. Code d'identification du créateur : **PF** (Patricia Forget)
5. Numéro de dossier : **0001**
6. Nom signifiant : **Recteur** (Nom d'une personne significative sur la photographie)
7. Numéro de chacune des photographies : **0001**

Toutefois, une dernière question demeure. À qui incombe ou incombera toute la responsabilité d'effectuer cette numérotation? À chacun des créateurs? Aux conservateurs? Ou à une personne spécifique du service qui, de par ses fonctions, assumerait également le transfert, l'identification, le rassemblement des pièces (images, vidéos, textes, etc.)? Tout reste encore à être défini.

4. Agir

Les personnes responsables des projets du 30^e me permirent d'engager des étudiants(es) pour effectuer, sous ma supervision, le traitement physique et intellectuel des photographies analogiques et numériques. Étudiantes en design graphique de l'ÉMI, elles connaissaient déjà les ressources essentielles à mon projet et avaient déjà de l'expérience dans le traitement archivistique de photographies. En effet, avec la collaboration du directeur du département de l'ÉMI et l'apport de la technicienne des ateliers numériques, nous avons eu accès aux laboratoires des ateliers numériques pendant tout le projet et surtout pendant l'été où seuls quelques rares privilégiés avaient accès à l'ensemble du matériel habituellement réservé à l'ensemble des étudiants.

¹¹ J'ai établi cette procédure en me basant sur la section 4 d'un document auquel j'ai participé, dont voici la référence : CREPUQ : Sous-comité des archivistes (2009, juin). *Mesures transitoires et bonnes pratiques de gestion des documents numériques*, consultée le 20 juin 2011, http://www.crepuc.qc.ca/IMG/pdf/mesures_transitoires_bonnes_pratiques_GDN-10juin.pdf.

Avec la collaboration de l'équipe de la DCR, le corpus des photographies entreposées au sous-sol a été transféré dans un local spécialement consacré au 30^e. De plus, la direction de l'établissement, plus particulièrement le secrétaire général, alloua un budget spécial afin que l'on puisse se procurer le matériel approprié pour la conservation préventive. Le travail s'est concentré spécifiquement dans un premier temps sur le traitement du matériel photographique analogique afin de le rendre conforme aux normes archivistiques en vigueur et d'essayer de trouver un classement approprié qui pourrait nous servir pour faciliter le travail du numérique et de la sélection des items qui serviront aux projets de diffusion du 30^e anniversaire.¹²

4.1. Corpus analogique(s) (épreuves photographiques, négatifs et diapositives)

Cette première démarche a permis à mon équipe de faire une découverte des plus étonnantes. Pendant 25 ans, le corpus de photographies analogiques avait été classé, répertorié et numéroté par le même employé, l'agent d'information. Chaque photographie avait été correctement classée dans un dossier-événement. Les informations inscrites sur chaque dossier-événement concernaient avant tout l'identification du contenu de la photographie : l'événement, la date, le lieu, les gens impliqués et dans certains cas, le nom de l'auteur de la photographie.

Classement d'origine

88-PH-001-1

88 : Année de création de la photographie, impliquant 19XX.

PH : Photographie

001 : Le numéro de dossier (chronologique)

1 : Le numéro de la photo (souvent repris du numéro du négatif de la photo)

Chaque dossier comprenait :

1. Une pellicule de négatifs
2. Une planche contact réalisée à partir de la pellicule de négatifs
3. Des enveloppes contenant individuellement une partie des épreuves photographiques (développées à partir des négatifs sélectionnés)

Précisions qu'il s'agit des années 1971 à 1973 ainsi que de 1986 à 1999.

Lorsque l'agent d'information part à la retraite en 1999, personne ne prend le relais. C'est souvent ce qui arrive aux fonctions uniques d'un service pour lesquelles l'individu n'a pas eu le temps de transférer sa méthodologie et son savoir-faire. Par conséquent, à partir de 1999-2000, plus personne du service ne classera, ne répertoriera, n'identifiera, ni ne numérotera les photos créées et commandées par le Service. Ceci correspond aussi exactement aux changements technologiques dans le service : passage de l'analogique au numérique. Le changement a été si radical que lorsque mon équipe commença le traitement du corpus analogique de ce service, personne ne se rappelait de l'ancienne méthodologie créée et mise en application par l'agent d'information, ce qui a eu des conséquences majeures sur la préservation et la conservation de la mémoire institutionnelle de l'Université. En effet, cela a créé une discontinuité.

¹² Le comité organisateur du 30^e avait en tête plusieurs projets de diffusion, dont un projet d'exposition virtuelle, un projet d'exposition permanente, un livre commémoratif et un gala.

La mise en valeur du travail des artisans du passé peut-être un excellent moyen de bien célébrer une institution en reconnaissant l'excellence de leur apport à travers le temps et du même coup, de diminuer la cassure qui s'était créée. C'est sur cet aspect que mon équipe et moi-même avons insisté. En effet, le travail exemplaire de l'agent d'information nous permettait de rapidement traiter presque l'ensemble du matériel analogique afin de le rendre conforme à la conservation préventive puisqu'il était déjà classé, répertorié, identifié et numéroté, ce qui nous a permis d'effectuer un inventaire incluant les données de base facilitant d'autant plus la sélection pour les projets de diffusion. Avions-nous une photographie du recteur de 1986? Avec notre nouvel outil de recherche, la sélection du matériel à numériser devenait un jeu d'enfants.

Cette découverte étonnante nous permit de mieux cerner les enjeux auxquels nous étions confrontés dans l'immédiat. En effet, de par la nature de sa fonction, originalement, c'était l'agent d'information qui avait la responsabilité de regrouper images, vidéos et textes d'un événement, de les identifier, de les numéroté et de les classer par dossier-événement. Est-ce vraiment par hasard que l'individu effectuant cette fonction devenait responsable de cette tâche? Est-ce que cette fonction est la mieux adaptée pour effectuer cette tâche? Ou doit-on rendre responsable l'ensemble des usagers ou une autre fonction? La question est lancée.

4.2. Corpus numériques(s) (partie I)

Il en allait tout autrement du corpus numérique. Étant encore utilisé, mais n'ayant jamais été réellement classé, numéroté et documenté, nous ne pouvions poursuivre notre tâche sans l'aide d'un des principaux utilisateurs, en l'occurrence, la conceptrice graphique. Le traitement de ce corpus passait inévitablement par l'institutionnalisation des procédures. Pour ce qui est du corpus des photographies numériques, le travail fut beaucoup plus délicat, puisque le numérique est un document dont la lecture requiert un appareil informatique, et delà provient toute la problématique. Ce qui peut rendre difficile le travail lorsque l'outil de lecture ne permet pas de rendre compte de l'entièreté des métadonnées intrinsèques à une photographie et selon les standards établis.¹³

5. Soutenir

Cette étude de cas m'a amenée à me questionner en tant qu'archiviste et muséologue sur les moyens les plus efficaces, judicieux, économiques et accessibles afin d'assurer la pérennité du patrimoine numérique, afin de soutenir tous ces comités dans la poursuite de leur but et surtout afin de mettre en place une plateforme institutionnelle permettant la gestion efficace du patrimoine photographique. La complexité des nouvelles technologies implique que l'ensemble du corps professionnel d'une institution s'y penche : informaticiens, archivistes, avocats, agents de communication, etc. Par conséquent, l'approche multidisciplinaire (muséologique) s'avère souvent la plus adaptée, puisqu'elle implique de considérer l'entièreté de la problématique afin de répondre à la mouvance de ces nouveaux environnements.

Défi majeur d'abord par la constitution même de nos administrations qui fonctionnent souvent en vase clos. Une volonté nouvelle de la classe dirigeante doit émerger afin de les conscientiser sur leur rôle,

¹³ Nous parlons standard de l'IPTC (International Press and Telecommunications Council) qui sont des métadonnées ajoutées par l'humain et de l'EXIF (EXchangeable Image File) qui sont des métadonnées fournies automatiquement par un appareil numérique.

mais également sur les solutions envisagées. Les dirigeants doivent non seulement permettre, mais favoriser la création d'équipes interdisciplinaires. Cette convergence d'un événement de célébration s'avérerait le catalyseur parfait pour permettre ce nouveau type de collaboration à l'Université. En fait, c'est ce que j'espérais. Entre la théorie et la pratique, comment arriver à bien faire adhérer les principaux usagers à votre cause? En fait, plus la solution réussit à répondre à une multitude de défis (technologiques, juridiques, économiques, etc.), touchant plusieurs intervenants, plus elle intéressera différents acteurs dans son développement. Par conséquent, plus elle répondra à une majorité de besoins, plus cette solution s'avérera gagnante.

C'est ainsi que j'entrepris d'abord des réunions interdisciplinaires avec les spécialistes des communications, des technologies de l'information (informaticiens et webmestre) et de la conservation, dans le but justement de trouver une solution qui répondrait à l'ensemble des besoins. L'ensemble des intervenants était d'accord pour acquérir un outil, un nouveau média, efficace pour l'acquisition, le traitement, l'accessibilité et la diffusion. Une solution qui pourrait faire d'une pierre trois coups, c'est-à-dire être utilisée pour entreposer de façon centralisée le patrimoine numérique (images, documents vidéos et d'autres « créatives files ») afin de l'organiser, l'archiver, y accéder rapidement, un DAMS (Digital Asset Management Systems), et ensuite la diffuser à une multitude d'utilisateurs internes et externes via les médiums du web. Nous en sommes même arrivés à sélectionner un outil qui permettait de répondre aux besoins de documentations, d'accès et de conservation à long-terme du patrimoine. Sans le support de la direction de l'Université, nous n'aurions pas eu le budget pour acheter cet outil.¹⁴

6. Institutionnaliser

Mais comment y arriver? En faisant accepter la responsabilité à chacun des acteurs avant même d'intégrer la solution. Les changements continuels de personnels permettent difficilement de responsabiliser les principaux acteurs ou peuvent au contraire devenir une possibilité d'amorcer un travail qui aurait dû être effectué depuis belle lurette. Depuis le départ de l'agent d'information, personne n'avait repris la tâche de classer, de répertorier, d'identifier et de numéroté les photos. La personne assignée au repérage de ce patrimoine, en l'occurrence le concepteur graphique, venait juste de quitter l'institution après plus de dix ans. Le repérage s'effectuait en vertu de sa mémoire personnelle. Comment s'assurer que la version conservée est l'authentique, l'officielle, la plus à jour, la mieux documentée sans la participation indéfectible des créateurs de cette information, sans la participation des principaux usagers?

De par la nature même du numérique, les possibilités de copies sont infinies ou presque. Contrairement à l'analogique, la représentation numérique d'origine permet la création d'une copie sans altération aucune. Encore faut-il avoir les appareils permettant ce niveau de copie.¹⁵ De plus, rien ne ressemble davantage à un fichier qu'un autre fichier. S'il n'a pas été au préalable identifié (numéroté et documenté), rassemblé en sous-ensembles, la tâche devient colossale pour quiconque s'y attaque, même pour ses photographies personnelles. Alors, imaginez dix ans de photographies dans une institution

¹⁴ Extensis Portfolio, consulté le 14 août 2012, <http://www.extensis.com/en/about/index.jsp>.

¹⁵ Il n'y a pas nécessairement d'altération du produit de départ lors de la création d'une copie. Encore faut-il être capable d'effectuer une copie dans le même format et au même niveau de pixels, et cela sans perdre les métadonnées intrinsèques d'origine. Par exemple, dans le cas qui nous préoccupe, les RAW des photographes peuvent avoir été copié sur un appareil incapable de supporter un tel format ce qui ne laisse d'autres choix à l'utilisateur que de créer un autre type de copie que celui de l'original et par le fait même une perte de qualité peut s'opérer.

universitaire : voilà le défi qui attendait mon équipe. La préservation du patrimoine numérique doit avant tout être validée et documentée par les responsables mêmes de cette information. Voilà pourquoi le cœur du problème réside dans l'inclusion des créateurs et des principaux utilisateurs au sein du projet de préservation. Alors qu'auparavant, nous pouvions nous en passer, il faut maintenant les convaincre d'y participer activement.

L'entrée en fonction de la nouvelle conceptrice graphique en 2010 ouvrit une brèche dans l'inertie généralisée qui sévissait depuis près d'une décennie. En effet, la nouvelle personne engagée n'arrivait plus à rien retrouver. N'ayant pas accès à la mémoire personnelle de l'ancien concepteur graphique, il devenait difficile, voire impossible de répondre adéquatement et rapidement aux demandes de ses collègues, d'autres services et des départements. Il est donc facile de comprendre ce qui l'a poussée à collaborer avec nous. C'est réellement avec son analyse du corpus, son aide et sa collaboration que nous avons pu arriver à déterminer un corpus de base à traiter au niveau des photographies numériques. Après collaborations et concertations, nous avons établi une arborescence où l'ensemble des photographies serait regroupé en sous-ensembles logiques selon les événements. La conceptrice graphique effectuerait un tri (élimination des doubles), le classement et le regroupement du corpus qu'elle utilise dans son ordinateur, tandis que les étudiantes s'occupèrent du corpus provenant des CD(s) et DVD(s) transférés, pour la plupart, par les photographes contractuels engagés par l'Université, ainsi que certains employés. Afin de pouvoir effectuer un tri et un classement, l'ensemble des photographies numériques a été transféré sur un disque dur externe. Après avoir été bien classés selon l'arborescence, les fichiers ont été gravés en deux exemplaires sur des DVD(s) de qualité archivistique : une copie pour l'utilisation de la conceptrice graphique et une autre copie pour la conservation à long-terme aux Services des archives.¹⁶

Encore fallait-il remédier à la situation, s'assurer d'avoir une méthodologie d'identification qui soit appliquée. Oui pour la création, mais non pour la conservation de l'authenticité, l'assurance de l'originalité, la conservation de la documentation rattachée autant au niveau de la matérialité du média (le contenant) que l'information qui lui est rattachée (le contenu). Dans les deux cas, le créateur devient essentiel puisque le numérique est un format plus instable et pouvant être facilement copié en plusieurs exemplaires. La spécificité de ces nomenclatures et du traitement de celles-ci sur l'information constitue un travail difficile, complexe, mais nécessaire. De par sa nature, souvent une seule personne connaît l'entièreté de ce processus, et lorsque cette personne quitte son poste, elle part souvent avec la clé de voûte. Le vrai défi devient donc de bien définir la responsabilisation de chacun des acteurs. Encore là, lorsque l'on réussit à convaincre la direction de l'établissement, la direction du service, la personne chargée de cette information, encore faut-il trouver un moyen qui permettra de bien intégrer, et ce, de façon systémique ces images afin qu'elles puissent être réellement intégrées dans la mémoire institutionnelle.

À la fin du projet, le directeur du service décida de modifier la fonction de commis à l'éditique pour technicien en infographie. Encore là, un changement de garde était à prévoir. Dans la fonction de ce nouvel emploi était mentionné qu'il « assurerait la tenue à jour du système de classement des photos. » Rien n'a encore été entrepris, par manque de temps. C'est donc rester un « vœu pieux » pour l'instant.

¹⁶ CDFinder (for Mac), consulté le 14 août 2012, <http://www.cdfinder.de/>; CDWinder (for PC), consulté le 14 août 2012, <http://www.cdwinder.de/>; Extensis Portfolio, consulté le 14 août 2012, <http://www.extensis.com/en/about/index.jsp>.

7. Extérioriser

Sans ce travail de débroussaillage, il aurait été impossible de bien comprendre le corpus de photographies pour ensuite effectuer une sélection appropriée. Seuls les éléments (photographies et documents) sélectionnés ont été numérotés, documentés et gravés sur des DVD(s) de qualité archivistique et rendus accessibles via le nouvel outil qui permet d'une part de rendre ce patrimoine accessible à plusieurs utilisateurs, mais également de le diffuser de façon permanente à l'aide de galeries que nous pouvons créer et diffuser sur le site web.¹⁷ Cette opportunité permettait d'analyser la situation de près et de répondre aux besoins criants de l'Université en matière de conservation de son patrimoine photographique. Ce projet permet de rendre accessible un patrimoine oublié de l'Université en devenant une des bases incontournables pour enrichir visuellement les projets d'envergure du 30^e, tels que le livre commémoratif, l'exposition virtuelle et le gala.¹⁸

8. Conclusion

À court terme, rien de bien épeurant pour la mémoire institutionnelle; à moyen terme, il est encore possible de rétablir la brisure. Cependant, à long-terme, le chaos, l'incompréhension et surtout l'opacité grandissent.¹⁹ Cette constatation se révélera encore vraie à la fin du XX^e siècle puisqu'une certaine stabilité institutionnelle existait encore. Avec le début du XXI^e siècle, l'avènement des nouveaux médias combiné à une économie et un marché du travail de plus en plus instable, façonne une situation peu reluisante pour nos institutions : changement rapide de personnel, changement constant de technologie, changement de méthodologie, changement de services. Ce n'est pourtant pas une querelle entre l'ancienne et la nouvelle génération, à savoir si nous sommes pour ou contre les nouveaux médias. Il faut définir leurs forces et leurs faiblesses en fonction des besoins et des capacités de l'institution. Mais surtout, il ne faudrait jamais oublier que l'humain a une capacité limitée d'adaptation face aux changements.

Il serait erroné, voire même dangereux, de penser que l'on peut simplifier ou réduire l'ensemble de la documentation et de vouloir appliquer une solution unique à un problème si complexe. Cela entraînerait irrémédiablement une perte majeure des connaissances associées à ces méthodes de documentation et surtout au contenu de ces informations, surtout avec l'avènement de nouvelles technologies qui rend de

¹⁷ Vous en trouverez un exemple concret sur le site de l'exposition permanente.

Université du Québec en Outaouais (UQO) (2011, décembre), *Historique de l'UQO (Exposition virtuelle créée par Patricia Forget, archiviste universitaire, grâce au 30^e anniversaire de l'Université)*, consultée le 14 août 2012, <http://uqo.ca/historique/naissance-uqah/deuoq> et <http://media.uqo.ca/deuoq/>.

¹⁸ Livre du 30^e, Université du Québec en Outaouais (UQO), *Bâtisseurs d'avenir : Histoire d'une Université qui voit grand* (Gatineau, Québec : Coopérative de solidarité des Écrivains des Hautes Terres et UQO, 2011), 299 p.; Exposition virtuelle du 30^e, Université du Québec en Outaouais (UQO) (2011, décembre), *Historique de l'UQO (Exposition virtuelle créée par Patricia Forget, archiviste universitaire, grâce au 30^e anniversaire de l'Université)*, consultée le 14 août 2012, <http://uqo.ca/historique/>; Gala du 30^e. Université du Québec en Outaouais (UQO) (2011, mars). *Capsules vidéo diffusées lors du Gala des 30 ans de l'Université du Québec en Outaouais (UQO) à la Maison de la culture de Gatineau*, consultée le 14 août 2012, <http://www.youtube.com/watch?v=3dhnewZ0uOg>.

¹⁹ « De nombreux détenteurs d'expertise partent, et le peu de moyen et d'outils déployés pour "récupérer" leurs savoirs entraîne une réduction substantielle de la performance de ce qui, dans certains cas, peut remettre en question la continuité des activités. » Rivard et Roy, 67.

plus en plus floue la ligne entre les données essentielles au bon fonctionnement d'une institution, et les médias de masse associés à la diffusion de l'idéologie de la société de consommation.

L'idée de progrès qui nous pousse constamment à revoir nos méthodologies nous fait oublier souvent qu'un système de repérage est efficace uniquement lorsqu'il est commun à plusieurs personnes et surtout lorsqu'il est facile d'utilisation. Néanmoins, il demande une certaine rigueur, un certain temps et ne peut être laissé au hasard. C'est une tâche qui n'est souvent pas prise au sérieux dans nos institutions qui voient toujours de plus en plus à court terme. Cependant, avec l'arrivée des nouveaux médias, où les formats sont en constante évolution, cette rigueur est de plus en plus justifiée, même dans le court terme, puisque la création d'images, de sons, de vidéos n'a jamais été aussi exponentielle qu'elle ne l'est présentement. Quoi conserver? Quel format? Pour combien de temps? Ces questionnements ne peuvent plus être laissés au hasard très longtemps sans des conséquences graves sur la performance même de l'institution.²⁰

Toutes les étapes de la conservation numérique (agir, reconnaître, soutenir et affermir, institutionnaliser et extérioriser) seront à reprendre et à reprendre. Comme tout changement ou tout savoir, ce n'est qu'à force de passer par les mêmes étapes qu'enfin l'ensemble pourra être intégré par l'institution, mais également s'adapter aux nouvelles données dans un monde en continuel changement. C'est, je pense, le principal rôle de la conservation du numérique ou du conservateur du numérique.

Dans un monde où « le temps c'est de l'argent » et où le format des supports se modifie en un temps record, nous avons encore moins de jeu pour la préservation efficace du patrimoine que nous créons. D'où l'importance d'effectuer ce travail en amont du processus de préservation. Il s'avère souvent plus efficace de bien identifier le lieu stratégique de convergence d'un patrimoine afin de pouvoir trouver la personne qui sera motivée de par ses fonctions à devenir responsable de l'identification. Lorsqu'un employé crée ou reçoit dans la cadre de son travail une photographie, quelques minutes suffisent à identifier, numéroté, rassembler en contexte ce patrimoine. Cette mise en contexte capitale peut s'avérer complexe, même pour le créateur, quelques mois seulement après la réception. Imaginez alors le travail des services de mémoire ou des institutions de mémoire s'ils doivent effectuer ce travail après coup. Avec les budgets de plus en plus réduits, la production de plus en plus grande de ce type de format, et ce, vu la facilité avec laquelle ce patrimoine peut être créé, il faut absolument remédier à la situation.

Bibliographie

Livre(s)

Brosseau, Kathleen, Mylène Choquette et Louise Renaud. *Normes de numérisation de la Société du Musée canadien des civilisations*. Gatineau: Musée canadien des civilisations, 2006, 35 p.

Couture, Carol. *Les fonctions de l'archivistique contemporaine*. Collection Gestion De L'information. Sainte-Foy: Presses de l'Université du Québec, 1999, 559 p.

« Selon les études menées sur le sujet, le non-traitement de la masse documentaire en entreprise peut faire perdre 5 % du chiffre d'affaires de cette dernière. On estime également que les collaborateurs d'une entreprise consacrent 20 % de leur temps à rechercher des documents ou à en recréer qui existent déjà », indique Elie Choukroun, directeur marketing et communication de Ricoh France, société de services de gestion documentaire; C. Ageneau, « Le rocher de Sisyphe », *Le nouvel Économiste.fr*, n° 1 604 (2012) : 32, consultée le 21 août 2011, <http://www.lenouveleconomiste.fr/lesdossiers/archivage-la-gestion-documentaire-14105/>.

- Charbonneau, Normand, et Mario Robert. *La gestion des archives photographiques*. Collection Gestion De L'information. Sainte-Foy, Québec: Presses de l'Université du Québec, 2001, 306 p.
- Eames, Charles. *A Computer Perspective: Background to the Computer Age*. Cambridge: Harvard University Press, 1990, 174p.
- Foulonneau, Muriel, and Jenn Riley. *Metadata for Digital Resources: Implementation, Systems Design and Interoperability*. Chandos Information Professional Series. Oxford: Chandos, 2008, 203 p.
- Gagnon-Arguin, Louise, et Sabine Mas. *Typologie des dossiers des organisations : Analyse intégrée dans un contexte analogique et numérique*. Collection Gestion De L'information. Sainte-Foy, Québec: Presses de l'Université du Québec, 2011, 213 p.
- Gagnon-Arguin, Louise, et Sabine Mas. *Classification des documents numériques dans les organismes : Impact des pratiques classificatoires personnelles sur le repérage*. Collection Gestion De L'information. Sainte-Foy, Québec: Presses de l'Université du Québec, 2011, 191 p.
- Leary, William. *Le tri des photographies en archivistique : une étude du RAMP et principes directeurs*. Paris : Unesco, 1985, 118 p. (PG1-85/WS/10)
- Manovich, Lev. *Le langage des nouveaux médias*. Dijon : Les Presses du réel, 2010, 605 p.
- Rousseau, Jean-Yves, et Carol Couture. *Les fondements de la discipline archivistique*. Collection Gestion De L'information. Sainte-Foy, Québec: Presses de l'Université du Québec, 1994, 348 p.
- Vodden, Christy, and Ian Dyck. *A World Inside: A 150-Year History of the Canadian Museum of Civilization*. Gatineau, Québec: Canadian Museum of Civilization Corporation, 2006, 104 p.
- Université du Québec en Outaouais. *Bâtisseurs d'avenir : Histoire d'une Université qui voit grand*. Gatineau, Québec: Coopérative de solidarité des Écrits des Hautes Terres et UQO, 2011, 299 p.

Article(s)

- Ageneau, C. « Le rocher de Sisyphe. » *Le nouvel Économiste.fr*, n° 1 604 (2012): 31-34. Consultée le 21 août 2011. <http://www.lenouveleconomiste.fr/lesdossiers/archivage-la-gestion-documentaire-14105/>.
- Ballay, J.-F. « Paradoxes de la transmission et de l'apprentissage dans un monde radicalement incertain. » *Télescope* 16, n° 1 (hiver 2010): 1-20.
- Charbonneau, N. « Le tri des photographies. » *Archives* 30, n° 2 (1998-1999): 29-42.
- Gadoury, N. « L'évaluation des photographies en format numérique. » *Archives* 41, n° 1 (2009-2010): 31-42.
- Robert, M. « Les archives photographiques à la ville de Montréal, » *Archives* 28, n° 3/4 (1996-1997): 81-95.
- Rivard, L., et M.-C. Roy. « Un cycle de rétention des connaissances pour combattre l'amnésie organisationnelle. » *Télescope* 16, n° 1 (hiver 2010) : 67-81.

Site(s) web(s)

- Bibliothèque et Archives nationales du Québec (BAnQ) (2009, septembre). *La numérisation des documents administratifs : méthodes et recommandations*. Version revue et corrigée (2010, juin). Consultée le 15 juillet 2010. http://www.banq.qc.ca/documents/services/archivistique_ged/Numerisation_des_documents.pdf.

CREPUQ: Sous-comité des archivistes (2009, juin). *Mesures transitoires et bonnes pratiques de gestion des documents numériques*. Consultée le 20 juin 2011.

http://www.crepuq.qc.ca/IMG/pdf/mesures_transitoires__bonnes__pratiques_GDN-10juin.pdf.

Musée canadien des civilisations (MCC) (2006, avril). *Exposition : 150 ans de culture, de collections et de découvertes au Musée canadien des civilisations*. Version revue et corrigée (2007, avril).

Consultée le 14 août 2012. http://www.civilisations.ca/cmc/exhibitions/cmc/150/m150_01f.shtml.

Université du Québec en Outaouais (UQO) (2011, décembre). *Historique de l'UQO (Exposition virtuelle créée par Patricia Forget, archiviste universitaire, grâce au 30^e anniversaire de l'Université)*.

Consultée le 14 août 2012. <http://uqo.ca/historique/>.

The Convergence of Cultural Heritage

Practical Experiments and Lessons Learned

Jeannette A. Bastian and Ross Harvey

Graduate School of Library and Information Science, Simmons College, USA

Abstract

While we typically think that the major challenges for digitization and digital preservation are technological, cultural and professional issues often predominate and are more difficult to overcome. Research in the convergence of cultural heritage institutions conducted at Simmons College, Boston, since 2009 reinforces this observation. This research is described, focusing on the Simmons Digital Curriculum Laboratory, digital convergence projects negotiated and completed by faculty and students, and lessons learned. Problems encountered, including concerns about controlling public access, communication difficulties, key personnel at sites being laid off, lack of technical skills, lack of strategic vision, and constraints of organizational hierarchies, are outlined. Recommendations are made.

Authors

Jeannette A. Bastian is professor at the Graduate School of Library and Information Science, Simmons College, Boston, where she directs their archives education program. She was Territorial Librarian of the United States Virgin Islands 1987 to 1998 and received her Ph.D. from the University of Pittsburgh in 1999. Her archival publications include *Owning Memory, How a Caribbean Community Lost Its Archives and Found Its History* (2003), *Archival Internships* (2008), and *Community Archives, The Shaping of Memory* (2009). For the past three years she has been part of a team initiating a Cultural Heritage curriculum at Simmons.

Ross Harvey is on the faculty of the Graduate School of Library and Information Science, Simmons College, Boston. Before joining Simmons in 2008 he held positions at universities in Australia, Singapore, and New Zealand. His current research and teaching interests focus on the stewardship of digital materials in libraries and archives, particularly on its preservation, and on the history of the book. His most recent publications are *Digital Curation* (Neal-Schuman, 2010) and *Preserving Digital Materials*, 2nd ed. (De Gruyter, 2011).

1. Introduction

For the past three years, faculty at the Graduate School of Library and Information Science at Simmons College in Boston have been engaged in conducting research and experiments in the digital convergence of cultural heritage institutions. Anticipating that most of our challenges would be technical ones, we were surprised to discover that our major issues were not technical at all but rather cultural, professional and organizational. In this presentation we bring these issues to you for your consideration and discussion by describing our research and experiments, the obstacles we encountered and the successes we achieved. We conclude with an analysis of the challenges of digital convergence based on our experiences.

Our research was two-pronged. On the one hand it was driven by a pedagogical interest in identifying and experimenting with a theoretical infrastructure that would accommodate digital convergence; on the other hand it was driven by a need to better understand the dichotomy between the current technical ability to achieve digital convergence and the many stumbling blocks and challenges encountered, particularly by small and medium-sized cultural heritage institutions.

Although the disciplines of libraries, archives and museums—philosophically, intellectually and often even physically linked for decades—have long been grappling with the chaotic and complex realities of connecting and converging, today digital technologies can finally turn visions of connecting and converging into reality. But despite the fact that recent research has produced a hefty body of digital convergence theory and that a growing number of major public and private institutions are committing to implementing convergence models, in practice the many obstacles to convergence are often overwhelming. In 2011, the National Library and the National Archives of the Netherlands announced plans to integrate into one organization, thereby following the paths of national libraries and archives in Canada, New Zealand and Ireland. This formal consolidation acknowledges the less formalized truths that library, archival and museum materials have always co-existed in all these domains (although, generally, they have only co-existed rather than merged or, in some cases, even communicated with one another). These institutions are committed to overcoming the many challenges to convergence, but for less highly visible and resourced cultural heritage institutions with far less leverage than these national institutions—though with equally critical public mandates for access—the challenges are formidable and sometimes almost impossible to overcome.

2. Convergence

Much has been written about the convergence of cultural heritage institutions over the past decade. A leading theorist in this area writes: “While the traditions and historical areas of expertise in archives, libraries, and museums may differ, the new challenges facing all collecting cultural institutions are best addressed in concert, in an inter-disciplinary forum that explores multiple solutions and takes advantage of many skills.”¹ Collaboration is an essential element in achieving convergence. Practitioners and theorists alike agree that “incorporating collaboration into the underlying work culture is foundational to realizing that institution’s potential and achieving its mission.”² But while collaboration may be an essential ingredient in the digital convergence recipe, other elements are equally crucial. Not only do these include a strong understanding of the potential of technology and digital curation, but they equally embrace leadership and organizational skills, as well as an appreciation of the theoretical constructs underlying the separate but related disciplines of archives, libraries and museums.

In our research at Simmons College we defined convergence in terms of the institution: a converging cultural heritage institution is one that combines library, archival and museum material, and is working towards a set of standards and best practices that unites traditional theory and operations from each. There are many reasons why libraries, archives and museums are finding the concept of convergence attractive. The potential for single searching across collections, the economic leverage of joint infrastructure investments for both physical and digital content creation as well as digital asset management, working in a common collaborative context, sharing similar organizational concerns, and the opportunity to identify gaps in the collections are just a few of the advantages. An overarching public mandate for seamless access and the potential of added user value through enhancements and new combinations (such as “mashups”) are primary motivators.

¹ Jennifer Trant, “Emerging Convergence? Thoughts on Museums, Archives, Libraries and Professional Training,” *Museum Management and Curatorship* 24 (2009): 369-386, p. 377.

² Gunther Waibel, *Collaboration Contexts: Framing Local, Group and Global Solutions* (Dublin OH: OCLC Research, 2010), 4.

The driver for convergence is, we contend, the use of technology for the representation, documentation, archiving, preservation and communication of cultural heritage knowledge. The outcome is the creation of new relationships and new knowledge by bringing digital data sets representing social and cultural activity together in novel ways.

But while libraries, archives and museums share many common concerns, roles and missions, they also come from distinct and different traditions. Not the least of these differences is that of professional education. The education of librarians, archivists and museum professionals is typically undertaken in separate library science, public history or museum studies programs. While each profession has its own established standards in areas of professional practice such as metadata schemas, search and data management tools, policies on access, the differences in organizational and professional culture may also be profound. As these distinctions carry over into careers, they may also become obstacles to crossing disciplinary boundaries.

3. Cultural Heritage Informatics

To ground our experimentation, we sought a theoretical framework and pedagogy that offered students a broad vision of cultural heritage institutions, looking beyond the silos of traditional information practice towards the confluence of a wide variety of data in both virtual and physical forms. We found that theoretical and pedagogical foundation in the concept of cultural heritage informatics. Cultural heritage informatics, a relatively new discipline arising from convergence, emphasizes collecting, managing, supporting, reconciling, merging, and making accessible digital data across a broad spectrum of libraries, archives and museums. It can be particularly applicable and effective when these different entities reside within the same institution, but also surprisingly challenging. Cultural heritage informatics offers an overarching context for the seamless connecting and merging of a wide variety of materials within and across traditional cultural information institutions.

Cultural heritage informatics is a phrase that seems to have struggled over the past two decades to gain currency. It embraces digital convergence, cultural heritage information technology, digital curation and an entire range of practices and theories. The word informatics is increasingly used in combination with other disciplines to express this technology connection as in, for example, health informatics, archeological informatics, social informatics, and community informatics. Jennifer Trant and David Bearman were among the early users of the term, beginning in the early 1990s through their series of conferences, International Cultural Heritage Informatics Meetings (1991-2007) and the ongoing Museums and the Web beginning in 1997. On their Archives and Museum Informatics web site they define informatics as “the interdisciplinary study of information content, representation, technology and applications, and the methods and strategies by which information is used in organizations, networks, cultures and societies.”³

The working definition designed by the Simmons group that has guided their particular set of cultural heritage informatics experiments follows similar lines. It asserts that cultural heritage informatics generally refers to the intersection between computer science and cultural heritage, a partnership between technology and the legacies of the past as found in such cultural heritage institutions as libraries, archives

³ David Bearman and Jennifer Trant, *Cultural Heritage Informatics 1999: Selected Papers from an International Conference* (Pittsburgh, PA: Archives & Museum Informatics, 1999).

and museums. Cultural heritage informatics focuses on the use of technology for the representation, documentation, archiving, preservation and communication of cultural heritage knowledge.

More specifically, cultural heritage informatics refers both to the study and creation of added cultural value by the linking of disparate digital data sets, stored either locally or remotely according to accepted standards of description, arrangement, and metadata for archives, records management, museums or cultural materials. It encompasses the appraisal of data and data sets for enduring value in the context of archives or cultural heritage, and explores the creation of new relationships and new knowledge by bringing digital data sets representing social and cultural activity together in novel ways. It is also concerned with the policy, social, economic, organizational and legal issues of digital culture and heritage from the perspective of stakeholders such as heritage institutions and other cultural participants. The scope of cultural heritage informatics includes standards, metadata and every phase of the application of information technology, such as data capture/digitization, preservation, information/data processing, reconstruction, visualization, and documentation, as well as the dissemination of the output of these technical processes to cultural heritage communities and the general public.⁴

4. The Research

Since 2009 a research team at Simmons College, including faculty and students, has been working with partners in cultural heritage institutions throughout New England. The goal was two-fold: to design a cultural heritage informatics curriculum for students in the Graduate Library and Information Science program; and to provide experiential learning for the students by working with partners on actual convergence projects. In this way, curriculum combines classroom learning and experimentation in a laboratory setting with practical experience. The team has undertaken three main activities: designing appropriate coursework; constructing a virtual laboratory using open source applications as an experimentation space for digital projects (see <http://calliope.simmons.edu/dcl/>); and collaborating with six different cultural heritage institutions on digital convergence projects. Conclusions drawn from this collaboration form the basis of this paper.

5. The Digital Curriculum Laboratory

A Digital Curriculum Laboratory (DCL) anchors the entire convergence project. This virtual space provides integrated access to digital content, content management tools, standards and curriculum-based scenarios, and allows experimentation with a wide range of open source applications relevant to digital convergence. Developed by the Simmons faculty and supported by external funding,⁵ the DCL facilitates scenario-building, problem-solving, evaluation, and tool utilization by making it possible for students to apply and assess a variety of online archival and preservation procedures and techniques. The DCL was envisioned as an open access space containing a variety of digital content, providing an array of digital

⁴ We are indebted to our colleagues Terry Plum, Martha Mahard and Michele Cloonan for this extended collaborative definition of cultural heritage informatics.

⁵ In 2009, we received a grant from the National Historical Publications and Records Commission that directly supported the development of a virtual curriculum laboratory. Simultaneously we received an Institute of Museum and Library Services grant to develop a cultural heritage curriculum. A component of this grant supported laboratory development.

asset management systems for describing, preserving and managing this content, offering sets of descriptive, content, structure, and data value standards, and offering an evolving set of instructional learning modules and exercises to prepare students for today's professional environment in cultural heritage informatics. The DCL was also envisioned as a permanent curriculum tool to facilitate experimentation with digital materials in a variety of venues and contexts. It was expected that students would bring digital issues from our partner sites into the DCL where they could test possible solutions.

6. Our Partners

In selecting institutional partners, we had a wide variety of choices in terms of both size and focus of the institution. Cultural heritage institutions abound in New England, which comprises the six states on the northeast coast of the United States and was the site of the earliest American settlements in the 1600s. People in New England are proud of their heritage and even the smallest town generally boasts a local historical society with a mixture of artefacts, records, books and paintings, often located in a historic house with appropriate gardens and landscaping. The cultural heritage institutions in New England not only offer a laboratory for experiments in cultural heritage informatics, but also pose all the issues and challenges of digital convergence inherent in small (and often struggling) under-resourced institutions. At the same time, they are also the institutions whose relatively hidden collections have the greatest need for access and whose lack of support requires the greatest need for creativity and innovation.

The cultural heritage institutions with whom we collaborated demonstrate a wide variety of aims, activities, scope of collections and technological expertise, and a significant diversity in size. All of them, to varying extents, contained museum, library and archival materials and were actively creating digital assets. Teams of students worked on aspects of creating appropriate digital convergence models that would fit each individual institution. The six institutions we initially partnered with included one public library with a special collection, one historical society, three museums each with a special focus, and one organization that served as an umbrella body for historic houses throughout New England. Each of these organizations was well-established with collections that spanned many decades; each included library, museum, and archival materials that co-existed side-by-side; each generally organized these materials using separate systems; and importantly, each desired to improve access through digital convergence and was involved in creating digital assets.

7. What We Encountered

Three of the projects will be briefly described, with a focus on the sites, the specific digital experiment, the results, and the many and unexpected issues encountered. These included concerns about controlling public access, communication difficulties, key personnel at sites being laid off, lack of necessary technical skills, lack of strategic vision, and constraints imposed by existing and ingrained organizational hierarchies. These led us to recognize the critical importance of understanding the policy, social, economic, organizational, and legal issues of digital culture and heritage from the perspective of stakeholders such as heritage institutions and other cultural participants.

Our initial contention was that it would be technical issues that impeded convergence, especially the difficulty of interoperability among systems and metadata. For example, one of our partners is a library of a large museum, both well-resourced. This library was formed from the amalgamation of two libraries with long histories. We suspected that there would be issues to resolve based on the use of

different metadata sets stemming from different library-based systems and also museum systems, and also that there would be difficulties posed by lack of interoperability between museum and library catalog applications. Our suspicions were confirmed, but, in addition and probably of greater significance, was a cultural aspect—the museum doesn’t talk to the library, which it considers as a junior partner, and this stymied any significant action towards convergence.

Our major success story was the collaboration with the Gropius House at Historic New England. Historic New England defines itself as a “museum of cultural history that collects and preserves buildings, landscapes, and objects dating from the seventeenth century to the present and uses them to keep history alive and to help people develop a deeper understanding and enjoyment of New England life and appreciation for its preservation.”⁶ Centered in Boston, Historic New England operates thirty-nine historic houses scattered throughout New England. Among these is the Gropius House, built by architect and Bauhaus founder Walter Gropius in 1939 as his family home when he moved to Massachusetts to teach architecture at Harvard’s Graduate School of Design. The family lived in the house until Gropius’s death in 1969. Historic New England acquired the house and surrounding grounds in 1979. Preserved intact, the Gropius House includes furniture, artwork, books, clothing, and archival materials—all the impedimenta of family life. Gropius and his wife also designed the grounds, and the Japanese Garden and sculpture installations have been maintained as they left them.

In designing and executing a digital convergence project, the student team focused on one room in the house, the bedroom of daughter Ati Gropius. This room contained a wide variety of objects, art work, clothes, and books. By happy coincidence, Ati Gropius Johansen visited Historic New England in Boston while the students were working on their project. They were able to interview her and to incorporate the interview into their web exhibit of the room. Their convergence project involved scanning and describing all materials in the bedroom, linking items to descriptions and creating an online exhibit. Thanks to the interview with Ati, which had audio and video components they were also able to link objects to the interview in a dynamic fashion. The students added their descriptions of the objects to the Historic New England database using that organization’s prescribed metadata schema and, using the applications in the DCL, experimented with finding the best content management system for the objects in the web exhibit. They also experimented with digital mapping of the room within the wider context of the Gropius House. The final exhibit provides a model for Historic New England in working with its other sites. It was featured in their monthly newsletter and will be a permanent part of their web site.

This was a convergence success story for both the students and the site, where the exhibit seamlessly moves through different formats and media to deliver a single integrated information experience. No issues arose from the use of information technology at this site, which has an effective Systems Librarian/Archivist (the very job title illustrates convergence) who is a recent graduate from Simmons GSLIS and was very willing to work with our student team. Both the Systems Librarian / Archivist and the Senior Curator the students worked with had agreed on an integrated vision of how their collections should be described, displayed and accessed and Historic New England had the organizational infrastructure and will to implement the vision.

At the other end of the spectrum of success is a less well-resourced institution, a museum of American textiles, costume and material culture that combines museum, archives and library roles around the mission of “telling America’s story” through these materials. In spite of this avowed mission, the

⁶ “Our Mission,” approved September 24, 2008, <http://www.historicnewengland.org/about-us/mission-and-vision>.

museum web site presence is entirely separated from its library, which contains an internationally authoritative collection in a wide variety of formats documenting the textile and clothing industry.

We were aware from the start of our collaboration that information technology issues would be significant. This institution relies heavily on a proprietary application that, for primarily financial reasons, has not been updated for some years. Projects initially identified as appropriate for collaboration included planning the migration of a large collection of digitized images stored on obsolete technology and mapping out the metadata into standardized metadata schema. Other projects included utilizing the DCL to create an Omeka digital exhibit to exemplify how unique objects in the collection could be enhanced and displayed through the versatility of free and open source software.

But these information technology issues, challenging enough, were trumped by economic realities. Major staff layoffs took place while the student team and staff from the museum were finalizing the projects. The few remaining library and archives staff, accustomed to years of financial vagaries, depend on the museum's board to find a solution but seem not to feel that they have any control of that process. While the staff at the institution were eager to cooperate on a convergence project, the lack of funding and support for the library and archives, and lack of an overall vision for collaboration and convergence were obstacles that goodwill could not overcome. Outdated technology and lack of upgraded systems were challenges for students that could not be offset by a great willingness to help find solutions.

Lying somewhere in between on our success spectrum is a public library with a special collection. Founded in the late nineteenth century, this renowned colonial history collection begins with the first settlements in the 1600s, focusing on the revolutionary events that took place in this region. In particular, the collection boasts a rich collection of primary sources on the literary history of the area, including materials on the transcendentalists and their heirs. The initial convergence project was the linking of a series of fourteen paintings to related holdings within the collection. The students would scan the selected artwork, select an appropriate digital asset management system, after experimenting with several in the DCL, describe the artwork, enter descriptions into the selected system, and finally link these to related library sources, creating navigation between the exhibit and web site.

Although this project ran into technology issues similar to those experienced in several of the other projects, the overarching obstacles were the lack of communication about expectations between the site and the students and the deliberately narrow focus of the institution. For reasons related to their view of the context of the collection, the librarians in the special collection were more interested in attracting users to the physical site than to the virtual site. For this reason, they permitted only a limited amount of information on the web site and restricted linking. Additionally, although the students created a model for the project, they were unable to actually link to the site's web site due to technical incompatibilities that were not made clear at the time that the project was initiated. The students completed their tasks and created a prototype but were unable to make it into the dynamic web site they had envisioned. Although the perspective and expectations of the site were perfectly legitimate in terms of the collections, they had not been sufficiently communicated to the students. While a more limited convergence project could have been designed to meet the needs of this site, more discussion and analysis as well as a better articulation of the overall vision would have helped to make this project more successful.

8. Lessons

These three examples illustrate that, although many of the convergence problems centered on technical issues, importantly the larger, overarching obstacles were not technical. Rather, they stemmed from the

organizational culture at the sites, from competing interests between creators and preservers and between access and preservation philosophies. The critical non-technical factors we identified are: mismatch between expectations of partner sites on the one hand, and student teams and faculty on the other; inadequate communication; and insufficiently defined expectations. Specific examples of the problems encountered were: concerns about controlling public access; communication difficulties between students and site staff; key personnel at sites being laid off; lack of necessary technical skills at the sites; lack of strategic vision in some institutions; and constraints imposed by existing and ingrained organizational hierarchies.

It is worth commenting on some of the technical issues, although they are not the primary focus of this presentation. We were surprised by the lack of technical expertise at many of the sites. At one site, for example, the institution's accountant handled the information technology operation, apparently as an act of good will, providing trouble-shooting measures and occasional support. We observed a lack of understanding of and interest in open source software, some of which (Omeka, for example) is well supported and very effective. Many sites remained wedded to outdated software, posing a significant challenge not only to site staff but also to the students. File-naming conventions were generally haphazard and where they did exist were not adhered to. One site preferred HTML over Encoded Archival Description (EAD), which uses XML, for its finding aids, impairing future interoperability for the institution.

The larger overarching obstacles we encountered included the organizational culture at the sites. Constraints were imposed by existing and ingrained organizational hierarchies. There were few integrated data systems, and at the same time a resistance to change. This was exemplified at one site where the lack of collaboration between the constituent parts of the institution resulted in significant impediments to convergence and efficiency. Digital images were created and managed in at least four different sections, each of which used different workflows, file-naming conventions, metadata standards, hardware and software. Although the need was recognized to standardize and centralize its disparate and disorganized image collections for more functional use, especially by the staff, this had not translated into any action. As a consequence issues such as difficulties when trying to locate a digital image file wasted staff time because of the need to search multiple systems; effort was duplicated in redigitizing images that had already been digitized; and, because there was no institution-wide oversight on digitizing policy and workflow documentation, collection material was digitized to inadequate standards.

Another larger, overarching obstacle we encountered was the competing interests between creators and preservers, and between access and preservation philosophies. One site digitized collection materials and made them available primarily to encourage use of the physical collections and staff expertise. The view was that this collection serves as the heart of the historical community of the town and is not for external users; the interest and needs of the local community were of primary importance to the institution and local materials should be interpreted locally.

The critical non-technical factors identified are worth dwelling on. One illustration of the mismatch between expectations of partner sites on the one hand, and student teams and faculty on the other, was the inability of the technical infrastructure at several of the sites to support the scope of the proposed solutions. The lack of technology resources and expertise at the sites often meant that there was a knowledge gap, and it became apparent that in some instances the questions raised during collaboration with the partners were not understood. Although Simmons faculty attempted to clearly define expectations with all partner sites well before the students participated, differences in expectations persisted. Criteria for selecting sites were established at an early stage in this research. The sites needed to

be physically accessible to students; to combine library, archival and museum materials in some way; to be interested in pursuing digital convergence of their materials; and to be actively creating digital assets. We held a group meeting of site staff, faculty made initial site visits where student projects were identified, students made site visits with faculty, then student teams worked independently at the sites and also off-site. This was, apparently, not sufficient to clearly define expectations.

The mismatch of expectations was based in part on inadequate communication, the second critical non-technical factor we identified. This applies to communication between sites and students as well as internal communications within the sites. It is also worth noting that other factors such as expectations changing during the project also contributed to inadequate communication. Although discussions about technology and infrastructure were held with the partner sites at the time of the partner agreements, inconsistencies and gaps did not become apparent until after the students began working at the sites. One site commented that students appeared to be locked into the technology they know rather than what is needed by the site (although we observe that this comment also applies in the other direction) and that students don't have enough experience to explain what they want to do. At another site it quickly became apparent that one section of that site had not communicated what it was doing to another section, causing problems for students who needed to work with both sections to complete the project as initially defined. Other examples of inadequate communication within sites are the lack of documentation about procedures; for instance, one site had no documentation about the metadata standards it used, or about the servers available and what should be stored on each of them.

The third critical non-technical factor we identified was that expectations were not defined with sufficient clarity. This is related in part to the second factor identified, inadequate communication. Although discussions about technology and infrastructure were held with the partner sites at the time of the partner agreements, the inconsistencies and gaps did not become apparent until after the students began working at the sites. This was hampered because most sites did not have a workable vision that included convergence.

9. Recommendations Arising from the Lessons

What can be learned from these lessons? Although they are to some extent generic lessons that can apply to all endeavors, aspects of them that are specific to digital converged environments can be identified.

The overriding lesson from our cultural heritage informatics collaboration with partner sites is that developing clear expectations is crucial. This takes a lot of time, can be frustrating, and requires excellent communication on both sides. In the most successful project with the Gropius House, the staff at Historic New England was very willing to commit to significant time involvement. This meant that expectations were articulated clearly and thoroughly and that staff were available for consultation when students encountered problems. Another clear lesson is that sites need to be better aware of what is needed to function effectively in the digital environment, in terms of both technical and general requirements. As noted, we frequently observed that the understanding of information technology requirements and possibilities was minimal, resulting in unwillingness to consider new applications, different workflows, or even an upgrade of installed applications. Also as noted, most sites did not have a workable vision that included convergence. From these lessons we can identify some recommendations for similar projects in the future.

The first recommendation is to work with sites to develop their understanding of what is required to function effectively in a converging digital environment and to assist them in developing their abilities to

function effectively. This is a significant issue, of course, requiring ongoing action from educators, professional associations, and, indeed, learning by all professionals. Any future collaboration in this area will probably need to build in an educational component.

The second recommendation is to assist sites to develop clear and workable visions of convergence and of what they are aiming to do in a digital environment. This is likely to involve working with sites to identify goals that are consistent with their business use cases, develop strategies consistent with these goals, identify issues either favoring or impeding the implementation of these goals and their sustainability, recommend behavioral as well as structural modifications to improve end-user interaction, recommend systems that support the strategies defined, and suggest workflows that maintain consistent value and integrity for digital content. We will incorporate these processes into the Simmons curriculum in a new course being developed for the Cultural Heritage Curriculum, Digital Asset Management for Libraries, Archives and Museums (DAM for LAMS).

The third recommendation is to be very clear about expectations when negotiating specific projects with sites. We expect to develop a scope-of-work form to assist us in negotiating expectations at the start of the project, noting in detail what we want, what we will use, and what we need.

One common element in all of our collaborations was the use of the DCL, developed as part of this research (<http://calliope.simmons.edu/dcl/>) and described earlier in this paper. It was fully used by students to investigate possible applications that could be used at the partner sites. As a result of presentations and publications the DCL is currently of interest to an increasingly wide range of professionals. It is designed to be used as a sandbox in which to test applications and software tools for their applicability to a local institution. We suggest that it is a model worth investigating by local consortiums to build and use as a local tool for testing appropriate technical solutions.

10. Conclusion

The Call for Papers for this conference notes specifically that “digital continuity requires meeting technological, legal, economic, political and cultural challenges” and makes explicit some of the cultural and professional challenges: “lack of cooperation among Information Technology, legal, archival, library, museum and other professionals or institutions; organizational and institutional culture; competing interests between creators and preservers and between access and preservation philosophies, evolving skill sets, cultural sensitivity.” We encountered most of these. Despite what we initially considered to be adequate preliminary planning (environmental scans, site visits, profiling, identifying practicum projects), the implementation of our planning was not entirely successful. The real issues of convergence and digital continuity go beyond translating theory into practice, but also, and probably more significantly, call for the recognition and negotiation of the myriad issues and concerns of the cultural heritage institutions themselves. Lack of resources, compartmentalized and siloed mindsets, territoriality, are but a few of these issues and concerns.

Digitization and Digital Preservation
Experiences in a Developing Country
Perspective

The Conservation and Preservation of Heritage in the Caribbean

What Challenges Does Digitization Pose?

Elizabeth F. Watson

The University of the West Indies, Hill Campus

Abstract

Grounded in the concept of SIDS, this paper looks at the conservation and preservation of Caribbean heritage. The contribution that the MoW programme makes to the conservation and preservation of heritage is also examined. The paper also looks at a number of issues associated with the digitization of this heritage and encourages professionals and policy makers to ensure that steps are taken to conserve and preserve the region's heritage so that future generations will be able to see and experience that which contributed to making them who they are as a people.

Author

Elizabeth F. Watson is Campus Librarian for the Cave Hill Campus of The University of the West Indies, which is located in Barbados. She is a member of the UNESCO National Commission for Barbados, Chair of the Barbados MoW Committee and a member of the Marketing Committee for MoW. She has written extensively on audiovisual archiving and Barbadian popular music.

1. Introduction

Internationally, it is becoming increasingly accepted and evident that there is equity of importance in the heritage of all states—size, geography, location, politics or other constructs are in reality of little moment. Heritage represents, to current and succeeding generations, who their ancestors were, what they did, what they knew, how they lived and what they valued *inter alia*. Heritage provides backward and forward linkages between generations. Heritage has been described as the “fingerprint of generations” (Freedom Park).¹ Thus, the preservation and conservation of heritage is a solemn obligation that must be undertaken by each generation.

For Caribbean countries, many of which have only recently emerged from the bonds of colonialism, heritage conservation and preservation assume an even greater significance. During enslavement and prior to independence, in many of these states, determined steps were taken to eradicate, suppress or devalue the heritage of the “Other”. As a consequence, scant respect was paid to the heritage of the majority who lived in the space known as the Caribbean. This has led to considerable and irreplaceable loss of heritage in some cases while, in other instances, many containers of heritage are becoming increasingly fragile and/or threatened by a range of degrading factors—manmade, natural or artificial.

A post-colonial shift taking place at the level of national consciousness has been the growing awareness of the importance of heritage in the Caribbean. Determined steps have been and are now being taken to privilege Caribbean heritage which traces its origin to Africa and, where possible, pre-colonial times in the region. The Africa gaze is due to the fact that the majority of Caribbean people are descendants of persons who were brought to the region from that continent. In Caribbean spaces where

¹ Freedom Park: A Heritage Destination. *What Role Does Heritage Play with Freedom Park?*, accessed September 16, 2012, <http://preview.tinyurl.com/a88fpxr>.

other ethnicities exist in large numbers, steps are also in train to privilege their heritage. These measures serve to counteract the historical hegemony in the region of a European-based heritage, which was one of the consequences of colonialism.

In this paper, the words “culture” and “heritage” will be used interchangeably and will be taken to mean a combination of things, i.e., the beliefs and behavioural characteristics of a group as well as that which is handed down from generation to generation. The term “cultural heritage” will also be used in a similar fashion. While cultural heritage was not included in the original definition of Small Island Developing States (SIDS), since 1994 it has become one of the characteristics that is and can be used to distinguish states that share these features from other states and SIDS states from each other.

This paper looks at the conservation and preservation of heritage using the concept of SIDS as its frame. Preceding any discussion of the challenges that digital technologies pose as conservation and preservation methodologies an understanding of the paper’s historical-cultural landscape is important.



Figure 1. The Caribbean.²

This paper’s perspective is geographically grounded in English-speaking Caribbean SIDS states, however, much of what is explored is also the reality of SIDS states in other locations. This positioning is not amiss because an examination of map and list of these countries indicate that most SIDS states are located in the Caribbean. The Pacific region has the second highest number of SIDS states. The major differences between these two regions are the vastness of the Pacific Ocean compared with the much smaller Caribbean Sea and their cultural differences. While these regions may have had different colonisers, they share a history of colonisation and all that is associated with that experience. For the purposes of this paper no distinction will be made between formats that contain heritage and/or cultural information.

² Carib Seek, *Map of the Caribbean*, accessed September 15, 2012, <http://preview.tinyurl.com/a9krvq2>.

2. What is SIDS?

SIDS was formulated as a concept in 1992. The underlying philosophy of SIDS is that small, low-lying island states have characteristics which present particular challenges. While geography and environment are the leading factors of difference there are also a number of other shared features which make SIDS a discernible group. Although the term SIDS refers primarily to island states, there are some continental countries which exhibit many of the same characteristics found in island SIDS and so they are also classified as SIDS. The following are the major characteristics of SIDS states.

- Low lying
- Small in size, with growing populations
- Experience a shortage of well developed, highly diversified skills sets
- Limited resources
- Susceptible to natural disasters
- Impacted by a high incidence of man-made disasters
- Remotely located
- Vulnerable to external shocks
- High transportation, energy and communication costs
- Legal frameworks and administrative practices are not sufficiently developed to deal with pace of development and rapid changes
- Limited opportunities to create economies of scale
- Considerable dependence on international trade
- Absence of a robust national economy based on national resources and a well-developed manufacturing sector
- Cost of technology
- Weak national currencies
- Many were once colonies of a European power



Figure. 2. Location of SIDS countries (modified from Wikipedia).³

³ Wikipedia, "A map of the small island developing states," accessed September 15, 2012, http://en.wikipedia.org/wiki/Small_Island_Developing_States (modified version).

Caribbean	Pacific	Africa, Indian Ocean, Mediterranean and South China Sea (AIMS)
 Anguilla	 American Samoa	 Bahrain
 Antigua and Barbuda	 Cook Islands	 Cape Verde
 Aruba	 Federated States of Micronesia	 Comoros
 Bahamas	 Fiji	 Guinea-Bissau
 Barbados	 French Polynesia	 Maldives
 Belize	 Guam	 Mauritius
 British Virgin Islands	 Kiribati	 São Tomé and Príncipe
 Cuba	 Marshall Islands	 Seychelles
 Dominica	 Nauru	 Singapore
 Dominican Republic	 New Caledonia	
 Grenada	 Niue	
 Guyana	 Northern Mariana Islands	
 Haiti	 Palau	
 Jamaica	 Papua New Guinea	
 Montserrat	 Samoa	
 Netherlands Antilles	 Solomon Islands	
 Puerto Rico	 Timor-Leste	
 Saint Kitts and Nevis	 Tonga	
 Saint Lucia	 Tuvalu	
 Saint Vincent and the Grenadines	 Vanuatu	
 Suriname		
 Trinidad and Tobago		
 United States Virgin Islands		

Figure 3. List of SIDS countries based on location (modified from Wikipedia).⁴

SIDS countries span the globe but are mainly located in three specific regions: The Caribbean, The Pacific and AIMS (Asia, Indian Ocean, Mediterranean and South China Seas).

The geographical, environmental, socio-economic and technological factors *inter alia* which separate SIDS states from others are however not the only distinguishing factors. In 1994, at a Global Conference on the Sustainable Development of SIDS, which was held in Barbados, one of the major documents emanating from this event was the Declaration of Barbados (1994).⁵ The first affirmation of this Declaration reads in part:

The survival of [S]mall [I]sland [D]eveloping States is firmly rooted in their...cultural heritage...[it is one of] their most significant assets; ...⁶

⁴ Wikipedia, "Small Island Developing States. List of SIDS," accessed September 15, 2012, http://en.wikipedia.org/wiki/Small_Island_Developing_States (modified version).

⁵ *Declaration of Barbados* (Bridgetown: Government of Barbados, 1994), accessed September 5, 2012, <http://www.un-documents.net/barb-dec.htm>.

⁶ *Ibid.*, n.p.

This almost 20 year old affirmation has elevated the importance of heritage to being a critical part of the SIDS firmament. Culture is one of the major attributes that distinguishes SIDS from other states and heritage is one of the features that makes it possible to differentiate one SIDS state from another. Unfortunately, to date, much more emphasis has been placed on environmental and other issues within SIDS rather than those which are concerned with heritage. This paper looks at the heritage issues of SIDS with particular reference to its conservation and preservation. Some consideration will also be given to the contribution that UNESCO's Memory of the World programme (MoW) has made to assisting SIDS to conserve and preserve their cultural heritage.

3. SIDS and Heritage

In an earlier section of this paper some general issues associated with the impact of history and colonisation on the heritage of SIDS states were explored. A Barbadian experience provides a specific example. In the early 1630s (the island was colonized in 1627 by the British), a critical culture bearer that the enslaved brought from Africa, the drum, was banned as it was seen as an instrument that could incite unrest and enable the enslaved to communicate with each other without the knowledge of the colonisers. Thus, one of the major cultural items of the enslaved became a forbidden object from early in Barbados' history. Other aspects of the heritage of the enslaved were similarly treated, or, not given space or time to be practiced. The significance of the drum to Barbadian culture has now been recognized thus, often it is central to any cultural performance on the island today.

In the post-colonial era in most SIDS states there has been ongoing cultural renaissance whereby determined and dedicated efforts being taken to research, revive and retrieve the lost, forgotten or fragmented heritage of these states. A particular challenge for SIDS countries is that much of its heritage is dispersed. There are large collections of SIDS heritage that are scattered throughout repositories in European capitals and beyond as a consequence of the nationality of the colonizing country and how colonial records were administered, managed and maintained. Thus, regional researchers wishing to consult these documents are regularly forced to undertake long and costly overseas trips in order to consult the records of their national or regional heritage. While under some circumstances this cultural heritage can be repatriated, such acts usually come at a cost to the receiving country. Sometimes the asking price (if a sale is on the table) is beyond the potential of a SIDS state's ability to pay. On the other hand, that which has remained in its *loci* of origin has often not been kept in the best of conditions and so degradation is often evident.

The impact of man-made and natural disasters is real. Few SIDS states have the ability to provide safe and secure environments to house their cultural heritage. Thus, wars, hurricanes, tsunamis and other events often have a long lasting negative impact on the cultural heritage of SIDS states. Sometimes there is nothing left to conserve or preserve. In the Caribbean, over the years, several archives and libraries have experienced extensive damage and the loss of heritage items through hurricanes. While Mali is not a SIDS state, it is a developing country and the recent destruction of several tombs at Timbuktu is an illustration of modern day damage to a nation's cultural heritage. In this situation, religious intolerance is the cause of the destruction of these significant heritage signifiers.

Another issue that SIDS states face is cultural penetration or cultural appropriation by large states particularly via technology. These two factors, in some instances, have dramatically changed the nature and essence of the indigenous cultural heritage of SIDS states, sometimes making it difficult for the un- or under-exposed to discern the authentic from the "corrupted" version.

The inclusion of a steel pan register in many electronic keyboards is a case in point. The steel pan is the national instrument of Trinidad and Tobago. Fashioned out of oil drums discarded by American troops billeted in Trinidad during World War II, the steel pan is universally known as being the most significant musical instrument to have been created in the 20th century. The steel pan is also an early achievement in environmental protection and preservation. The Japanese, the initial innovators of electronic musical instruments, created and included steel pan registers in many of these instruments. While these registers are tonally and sonically different from the original instrument they enable keyboardists to simulate a steel pan sound in their performances. The inclusion of steel pan registers in electronic musical instruments represents the use of technology as a means of re/presenting a key signifier of the cultural heritage of a SIDS state. To the unattuned ear the sound from these registers may be acceptable—however, for those accustomed to the live sound of a steel pan there is a vast tonal and sonic difference between original pans and electronic simulations of this instrument. One of the factors contributing to the inclusion of steel pan registers in electronic musical instruments is that the U.S. Patent Office refused to grant a patent for the steel pan to Trinidad and Tobago. Hence, non-Caribbean entrepreneurs with access to considerable cash were able to appropriate and commodify this important signifier of Trinidad and Tobago's culture. While the steel pan is not a document, its story illustrates how the cultural heritage of SIDS states is under threat.

4. The Memory of the World Programme: A Sensitizing Development for the Conservation and Preservation of Caribbean Heritage

The underlying philosophy of the MoW programme is to protect the world's documentary heritage. From the programme's perspective "documents" has a very broad interpretation. Papyrus, clay, paper, fabric, vinyl, ferrous tape, nitrate film and any other tangible objects capable of storing information are interpreted as documents within the framework of the MoW programme. While documents are not the only sources of heritage, because the MoW programme is the oldest initiative of its kind within modern times, the programme has helped to sensitize practicing information professionals and other stakeholders in the area of heritage to the issues surrounding the conservation and preservation of cultural heritage. The wording of part of its founding documentation is key to this paper. Established in 1992 the pertinent words are as follows:

Documentary heritage reflects the diversity of languages, peoples and cultures. It is the mirror of the world and its memory. But this memory is fragile. Every day, irreplaceable parts of this memory disappear forever. UNESCO has launched the Memory of the World Programme to guard against collective amnesia calling upon the preservation of valuable archive holdings and library collections all over the world....⁷

The vision of the programme is simple but powerful: protection of the world's heritage as it is a mirror of the world and its memory. Theories of identity and memory are beyond the scope of this paper; however, the inclusion of the world "memory" from the onset of MoW's establishment is an indication of how critical heritage is to identity formation. There are no delimiters on who can make nominations to a MoW

⁷ UNESCO, Memory of the World Programme, n.d., http://portal.unesco.org/ci/en/ev.php-URL_ID=1538&URL_DO=DO_TOPIC&URL_SECTION=201.html, quoted in R. Harvey, "UNESCO'S Memory of the World Programme." *Library Trends* 56, no. 1 (2007): 259-274, p. 260.

register.⁸ Inscriptions on the registers—national, regional or international—are entirely based on nominations meeting the criteria of the programme and dossiers being presented in an acceptable format. Thus, small states such as SIDS countries have the same right to make nominations as do the large and powerful countries of the world. It can therefore be said that the MoW programme is very democratic, treating all states, institutions and individuals with an even hand.

Vannini provides a Latin American and Caribbean perspective on the MoW programme when she writes that the programme is “an international effort to safeguard the at risk documentary heritage to democratize its access and to raise awareness about its importance.”⁹ An examination of the MoW International Register reveals that the following Caribbean SIDS states have either single or joint nominations on this register: Bahamas, Barbados, Belize, Dominican Republic, Jamaica, Guyana, Netherlands Antilles, St. Kitts and Nevis, St. Lucia as well as Trinidad and Tobago. Trinidad and Tobago has six inscriptions on this register of which four are single and two are joint. Further scrutiny of MoW’s International Register indicates that, to date, the Caribbean has been the most active SIDS region in the programme. The Caribbean’s high level of involvement in MoW suggests that the region is taking active steps to identify its cultural heritage and that it is cognizant of its responsibility to guard its heritage for current and generations to come. The region’s involvement in the MoW programme also indicates that people in the Caribbean have begun to place a value on their heritage, which is the antithesis of what happened in the heritage of the region during colonialism.

One of the requirements of having an inscription on a MoW register is that the nominator has to indicate what conservation and preservation measures will be taken with regard to the collection being nominated. This obligation becomes a critical incentive to stimulating discussions and action in conservation and preservation at national levels. For the Caribbean therefore, MoW has served as an important modality to encourage action with regard to conserving and preserving regional heritage.

Two of the enduring outcomes of the MoW programme are that the programme has increased interest in the culture of the “Other”. It has also stimulated interest in visiting the location of the MoW inscription to view the inscribed documents. MoW inscriptions have become additional stimulants to cultural tourism, a sector in the tourist industry that many SIDS states are keen to develop. The MoW programme therefore provides opportunities for the cultural riches of all to be shared by all.

Although this paper is concerned with the documentary heritage cognizance is taken that within SIDS states often it is the intangible heritage that dominates in many of these societies given that they are oral in nature and practice. The ensuing discussion on conservation and preservation is of equal value to both the tangible (documentary) and intangible heritages.

Although the concept of SIDS and the MoW programme were developed based on different needs and within different communities, that they were created in the same year, 1992, is a happy coincidence. SIDS provides a platform for countries with challenges that were not previously recognized as being critical to sustainability and survival while the MoW programme was created to preserve the heritage of all states. Through the MoW programme the heritage of small states can be privileged in ways that were not possible prior to the establishment of this programme.

⁸ Individuals as well as private and public institutions are eligible to make inscription nominations to any of the three registers of MoW.

⁹ M. Vannini, “The Memory of the World Program in Latin America and the Caribbean,” *IFLA Journal* 30 (2004): 293.

5. Conserving and Preserving Caribbean Documentary Heritage: Challenges

In this section specific issues to do with conserving preserving the documentary heritage of the Caribbean will be explored. As a region with many SIDS states, there are several factors which impact on the preservation of the Caribbean's heritage. These include a shortage of persons with the appropriate expertise; the cost of technology; a lack of easy access to and the cost of required consumables; remoteness from sources of supply; and, a susceptibility to natural and man-made disasters. These are some of the characteristics, listed in an earlier section of this paper serve to distinguish SIDS states from groups of states. In addition, there are several particulars that are intrinsically associated with the region's documentary heritage.

Caribbean heritage is dispersed as a result of the region's colonial history. For some single states, this heritage is located in several European countries. The history of the Republic of Trinidad and Tobago is a useful case in point. Metropolitan countries holding records of this twin-island state are England, France, Holland, Latvia (Courlander during colonial times) and Spain as, at some point in time, these were the metropolitan countries that ruled one or more of these islands. Trinidad and Tobago only became a twin-island state on the attainment of independence in 1962. While some records are duplicated in Port-of-Spain, many originals (and often only copies) are located in external repositories—official and private. Repatriation is not often a solution because of cost and/or unwillingness of the holding institution to deplete their holdings. Further, within the Republic records are also dispersed because of the different governance structures that have existed over the years. This dispersal has also contributed to the degradation of some heritage containers. The scattering of this state's records makes it very difficult to have a full and true picture of the extent of records of Trinidad and Tobago.

The failure to provide finance at levels that facilitate the conservation and preservation of the cultural heritage in SIDS states is well known. While policy makers may empathise with and understand requests to conserve and preserve heritage, this does not translate into the necessary financing being made available to do what is required. On this Smith states:

...government officials are known to wax eloquently on the heritage of our past and its value to present and future generations. But in reality it has fallen to the private sector to act as stewards of large parts of our recorded past....¹⁰

For example, the general dearth of appropriate housing within the region for cultural heritage owned by the public sector negatively impacts on the physical condition of such items. To some extent the lack of financing for the conservation and preservation of heritage in SIDS societies is because policy makers are generally unaware of the national and international value of their heritage and what its conservation and preservation means to future generations. This disconnect is often the cause of under financing that leads to format degradation of various kinds. In official circles, competition for support between the aesthetic and social welfare are unrelenting. It is challenging to defend arguments about the social value of cultural heritage compared to societal needs for running water and roads. Decision makers, including some information professionals domiciled in SIDS states are not fully seized of how heritage can be leveraged to good advantage. While not writing specifically on the situation in SIDS, Smith's words nevertheless resonate with such spaces. She writes:

¹⁰ A. Smith, "Valuing Preservation," *Library Trends* 56, no. 1 (2007): 7.

However, most people . . . are not aware of the economic costs or societal benefits of providing access to cultural content over an extended period of time. This lack of understanding has proven to be a significant stumbling block in securing adequate resource to preserve our analog collections.¹¹

In the Caribbean, it is often the private sector that has taken the lead in conserving and preserving the region's heritage. This comes at a cost, however, because issues like the public good take second place to the mission of the private entity. Public access also is often at a price—whereas access to one's heritage should not always be accompanied by a price tag. A comparison between the work and activities of the Barbados Museum and Historical Society (BMHS) (a semi-private entity) and Barbados' Department of Archives (DoA) (a government department) indicates that the BMHS has made considerable strides in conserving and preserving the heritage of the island and region that it owns. Considerable effort has also been made to add value to the holdings of the BMHS. Conversely, little has been done by the DoA in terms of adding value to its holdings and moving the institution from being a storehouse operation to an archive in the modern meaning of such institutions.

Professionals are responsible for making the case to conserve and preserve a nation's heritage. Often, within this cohort, there is not always the proactive stance that one would wish. Smith advances that professionals need to put forward more compelling arguments as to why the cultural heritage needs to be conserved and preserved.¹² Smith also warns that the lack of use should not be taken to mean that the object has no value.¹³ The value of heritage accrues over time and by not taking the necessary conservation and preservation steps now, future generations will be denied access to their heritage, an essential contributor to their understanding of their past, present and who they are.

6. Digitization: A Conservation and Preservation Strategy

Advances in technology and communication have opened new conduits and formats for conserving, preserving and accessing cultural heritage. Foremost amongst them at this time are digitization, which provides images of the objects and the Internet, which facilitates remote access to these objects. These technologies have, as Smith opines, "accelerated the demand for access to information."¹⁴ Digitization and the Internet help to level the playing field between rich and poor, north and south, developed and developing countries in that they expedite unprecedented and fast access to content. In addition, these technologies provide the means of addressing interest in the culture of the "Other" and expose the heritage of SIDS countries in unparalleled ways. The challenge for SIDS states however is to develop the internal capacity to digitize their cultural heritage, acquire the required and appropriate technologies as well as to accrue the knowhow to keep such objects accessible at all times.

Despite the known challenges, it is becoming increasingly clear that if SIDS countries want to make strides in conserving and preserving their cultural heritage, digitization is the leading option. In addition to expanding access to content this technology also provides a platform through which content holders can financially exploit their holdings thereby creating a revenue stream that is not time sensitive. Such income can help to underwrite the costs of conservation and preservation. The commodification of

¹¹ Ibid., p. 5.

¹² Ibid., p. 7.

¹³ Ibid., p. 11.

¹⁴ Ibid., p. 7.

cultural heritage is also likely to contribute hard currency to the coffers of SIDS states which are usually strapped for such assets.

Jamaica's *Daily Gleaner* (*The Gleaner*) has been digitized from its first issue, which appeared in 1834. This fully searchable full page database provides access to a wealth of historical information (over a million pages) that is generally not available from any other source. *The Gleaner* has attractive subscription rates and has streamlined its accounting practices whereby access can be purchased on an anytime, anywhere basis for a day, three months or a year. This means that the company is earning foreign exchange on a 24/7 basis. Digitization therefore has become a revenue stream of foreign exchange for *The Gleaner*. In 2012, the digitization of *The Gleaner* assumed important significance as part of the research process to develop plans for the island's 50th anniversary of independence, which was granted in August 1962.

The Lascelles Family, headed by the Earl of Harewood named after the family seat in Harewood, Yorkshire, England, garnered their wealth over the period 1648-1975 through their extensive Caribbean landholdings—primarily in Barbados and Jamaica; trading in the enslaved; and, merchant banking. The papers relating to the activities of this family in the Caribbean during this critical period of the region's history are chiefly located within the walls of the family seat and the West Yorkshire Archives. In 2012, the Earl of Harewood presented digital copies of these holdings to The University of the West Indies' Cave Hill Campus in Barbados and to the BMHS. Through digitization it is now possible to access, in Barbados, records which were either only known about or are new to the research community on the island.

These examples indicate how digitization has helped to conserve and make accessible the documentary heritage of the Caribbean. A major challenge in the digitization of heritage is that as technology advances digital platforms change, upgrade and/or become obsolete. Strategies therefore have to be put in place to ensure that digital objects remain robust and accessible. Two issues need to be addressed:

1. How should originals be treated? and
2. What digital preservation strategies will be used?

While in some quarters there is the feeling that originals can be discarded, in reality this is a dangerous policy. If digital surrogates become corrupted or inaccessible for any reason, the existence of the original makes it possible to recapture digitally the inaccessible digital content so that it is once more available. Digitization is therefore, contrary to the belief in some quarters, not a substitute; rather it is an alternative option to provide access. It is also a methodology that assists in the preservation of originals as digitization reduces the amount of physical contact the originals experience. Therefore, it is absolutely crucial for SIDS states to ensure that their heritage is kept in its original state and that arrangements are made for their safe keeping in proper conditions. In addition, library and archive collections accrue value based on their holdings of original documents and not digital surrogates.

There are now on the market many products being offered as digital preservation solutions. However, many of the challenges associated with digital conservation also impact on the preservation of heritage in SIDS states. The lack of expertise and the lack of funding are the two leading issues. Many professionals in SIDS states feel that once an object has been conserved and digitized that nothing else needs to be done. One of the major mistakes frequently made is that no provision is made for storing backup copies of the digital surrogate in a safe secure off-site location. Off-site storage ensures that if anything should happen to the institution or digital objects that it holds, it will be possible repopulate the

digital platform with its pre held content. Another omission/oversight is that no consideration is given to how digital platforms will be updated and made current with changes in technology—including the unfortunate experience of the provider of the preservation system going out of business! This issue is of critical import to the digital preservation process. As digital platforms are usually propriety technology the conversion of one's holdings from one system to another is fraught with technological difficulties and challenges. The choice of a preservation system therefore has to be carefully considered with much thought being given to what is possible in the short and long-terms.

Given the small size, financial challenges and dearth of expertise in SIDS states, one approach that heritage institutions in these countries could consider is standardizing conservation and preservation strategies in order to achieve economies of scale. It is recognized, however, that there are several internal issues that would have to be skillfully negotiated in order to counteract known and real challenges. These include that heritage institutions often belong to more than one entity. Within the cohort of owners there would exist different reporting structures, different sources of funding and different missions. Any cooperative approach would have to be expeditiously handled and accompanied by a willingness to see beyond the “I” to the “we” in order to achieve what is best for the nation as a whole. The provision of solutions to these challenges is beyond the scope of this paper but they are recognized as being real and in need of serious reflection.

7. Conclusion

Grounded in the concept of SIDS, and using the Caribbean as its geographical focus, this paper has explored a number of issues that impact the conservation and preservation of the heritage of such states. The paper also provided a number of examples and historical reasons why the heritage of the region is diffuse and not located in one space or place. The issue of repatriation was also considered as were the several factors that stood as obstacles to such movement.

The possibilities of using digitization for conservation and preservation were explored, as were some of the challenges that SIDS countries face in any attempt to digitize heritage. While there have been several positive steps towards the conservation and preservation of the heritage of SIDS states much more needs to be done. Professionals charged with managing these assets have to be in the vanguard of initiatives required to advance heritage conservation and preservation in SIDS countries. The aims of the MoW programme can serve as a very valuable tool to support any arguments being advanced to conserve and preserve heritage in SIDS states—most of which are members of UNESCO. Given the fragility of many heritage containers in SIDS countries it is imperative that conservation and preservation professionals in these states act without delay. Such action becomes all the more pressing because as Smith rightly argues:

A people who do not own and control their own cultural heritage are a people who can be held captive by false histories, fabrications and lies.¹⁵

¹⁵ A. Smith, “Valuing Preservation,” *Library Trends* 56, no. 1 (2007): 4-25.

Digital Preservation of Demographic Heritage

Population Censuses and Experiences in Mali and the Democratic Republic of the Congo

Richard Marcoux, Laurent Richard and Mamadou Kani Konaté

Abstract

To preserve and promote the demographic heritage of French-speaking States in some African countries is one of the goals pursued by the Observatoire démographique et statistique de l'espace francophone (ODSEF) at Université Laval (Quebec). The demographic heritage concept opens new ways on how to think about microdata collected by census operations that goes far beyond the statistical analysis challenges. This paper presents the main experiences that led to the creation of two digitization workshops of census paper schedules; the first one in Bamako (Mali) and the second one in Kinshasa (Democratic Republic of Congo).

Authors

Richard Marcoux is a full professor in the department of sociology at Université Laval and director of the Observatoire démographique et statistique de l'espace francophone (ODSEF), (*Demographic and statistical research institute for French-speaking countries*). He has addressed conferences and published numerous works in scientific journals and reviews covering international demographic issues, with particular focus on Africa and Quebec. In 2009, he directed a collaborative work published by Presses de l'Université Laval, titled "Mémoire et démographie : regards croisés au Nord et au Sud."

Laurent Richard is a research professional at the Observatoire démographique et statistique de l'espace francophone (ODSEF) at Université Laval and an analyst at Statistics Canada. Apart from helping to organize ODSEF digitization workshops in African countries, he has also been one of the coordinators of the Canadian Century Research Infrastructure project (CCRI).

Mamadou Kani Konaté is a sociologist and coordinator of ODSEF activities in Africa. For more than 20 years, he has published numerous scientific works on population issues and has been involved in noteworthy international partnerships, in particular with statistics institutions in Africa. He holds a master's degree from the Ethno-sociology Institute of the University of Abidjan and a Diplôme d'Études Approfondies (post-graduate diploma) in rural sociology from the Paul Valéry University - Montpellier III.

ODSEF¹

The Observatoire démographique et statistique de l'espace francophone (ODSEF) was set up in Quebec in the spring of 2009 following the signing of a protocol of understanding between the Government of Quebec, the Organisation internationale de la Francophonie (OIF), the Agence universitaire de la Francophonie (AUF) and Université Laval. Two objectives define the two-pronged orientation of ODSEF activities. The first objective involves taking steps both to preserve and promote the demographic heritage of French-speaking States, at a time when the heritage of certain African countries faces serious threats, so justifying the need for urgent, decisive action. The second involves supporting all initiatives seeking to

¹ www.odsef.fss.ulaval.ca

circumscribe the linguistic dynamics so as to get a deeper understanding of the place occupied by the French language, not only within French-speaking populations, but elsewhere as well.

In the last two years, ODSEF has undertaken a number of investigative missions to assess the state of demographic heritage in Africa, and has worked in particular with statistics institutions responsible for census-taking in Mali, Senegal, Benin, Burkina Faso, Niger and the Democratic Republic of the Congo. With financial support from the Institut francophone numérique (IFN), ODSEF supervised a pilot project to set up a digitization workshop at the Mali institute of statistics (INSTAT) in Bamako. In less than one year, the project compiled an electronic record of more than one million documents from the 1976 census. The success of this pilot project helped ODSEF obtain further funding from the IFN to continue the Mali project for the 1987 and 1998 censuses and set up a similar workshop in Kinshasa so as to record the only census taken in independent Congo, that of 1984.

Aside from the preservation activity, ODSEF also supports initiatives to gain recognition for census data and hosted some thirty researchers from fifteen African countries to work on data from their different countries. A number of young researchers who stayed at Université Laval now use their investigations to take part in the second ODSEF objective which is to support all initiatives geared towards developing census data.

1. Introduction

Significant advancements have been made in demography and other scientific disciplines in recent decades and this has been made possible by innovative breakthroughs in methodology and technology. It is worthy of note that it has become increasingly easy to acquire and process recent demographic and statistical information. Certain national institutions, like Statistics Canada, allow access to recent census and survey results and even to samples of raw data just two to three years after the data collection operation. Initiative de démocratisation des données (*Data Liberation Initiative*) was a major program that, to a certain extent, revolutionized the practices regarding access and processing of data collection operations at Statistics Canada (Moon, 2000; Poirier and Le Bourdais, 2009). For southern hemisphere countries, Macro International, leader of the vast worldwide program of demographic and health surveys (EDS), has set up a search engine and other tools that help provide access to data from numerous collection operations conducted in Latin America, Asia and of course, Africa.

Furthermore, widespread availability of access to more powerful and more advanced computer equipment has greatly contributed to making it easier to process data, for this is an important element in the work done by demographers. Our tiny portable computers, lighter and more powerful than ever before, now give us the capability of analysing databases for more than 15 to 30 million persons, equivalent to the total census data for the population of several countries. That was an impossible task at the end of the last century, less than 12 years ago!

Notwithstanding all this progress, two pernicious effects of these technological advancements can be noted. Firstly, a frenzied rush is in progress to obtain new data and to access the very latest information. There is now a very high demand for up-to-date indicators and the search for the information needed to construct them is often done to the detriment of any real thought to the social and demographic processes that influence these very indicators. The result is that older databases are often underused, a fact that weakens our comprehension of demographic processes which for the most part play out over the medium or even long-term.

The second factor that can be considered as having a negative effect is the widening gap separating researchers in the south and in the north in their ability to access demographic information, especially in French-speaking countries. Researchers in Quebec, Belgium, France and Switzerland for example, now have fairly open access to an array of chronological data about their different countries which, despite the changes in definitions and concepts, help to define a certain number of trends. Numerous scientific collaborations between demographers, historians, archivists and statistics administrators have proven to be very helpful in performing new evaluation of old surveys and censuses. Many of the initiatives currently underway made it possible to question certain myths and prejudices regarding the demographic past of the populations in these countries. At the same time, and in contrast with the above, countries in the south and in French-speaking Africa especially, often have great difficulty finding traces of censuses, civil status or even surveys that preceded recent data collection operations. This therefore limits the abilities of scientists to effectively delimit and understand the demographic transformations that they set out to study.

2. Demographic heritage

Just what is meant by “demographic heritage”? Most dictionaries define the term “heritage” more or less as follows: “Property that is inherited from past generations,” “property, valued possessions passed down collectively by ancestors,” “common heritage of a group.” The word “demography” comes from the Greek *graphè*, the “action of describing,” and *demos*, which means “people,” “population”. As such, demographic heritage refers to all sources that make it possible, or have made it possible in the past, to describe populations. Demographic sources are often the only written traces that provide a window onto the many social and economic characteristics as they relate to individuals who make up the population of a given territory. Taken together, this data forms the demographic memory of the individuals in a collective group, and as such, represents a high cultural value.

Although they can take different forms, demographic sources are generally divided into three main types: 1) registers and administrative files (civil status, parochial registers, etc.); 2) population censuses; 3) surveys. Registers and censuses represent the only exhaustive form of data compilation since they list each person in the population by family name and first name. In contrast, surveys include samples that are sometimes quite small. The statistical representativeness of a survey is generally based on census data (survey base) so as to be able to deliver “statistically significant” results, even though they may rely on samples often representing less than 1 % de la population. The fact that these surveys are not exhaustive diminishes their patrimonial nature accordingly, in comparison with all the demographic sources together. On the other hand, civil status registers are usually very specialized and cover a single theme in detail (births, marriages, deaths, etc.). They can be cross-referenced with other sources of information so as to facilitate more in-depth analysis.

Despite the criticisms they receive, population censuses are often the only source of quality information that covers the entire population. Civil status registers are usually unequal in quality depending on the sub-populations living in a country. In several regions of the world, the civil status contains gaps that make it unusable for other purposes other than administrative work. This is essentially the case for most countries in Sub-Saharan Africa (Dackam, 2003). In the 1970s, a United Nations study estimated that the rate of coverage of populations relying on civil status systems reached 99% in Europe and North America, whereas they scarcely reached 26 % in Africa (Laboratories for Population Statistics, 1976). Information available for the most recent period suggests that the civil status coverage in Africa

has scarcely improved over the last 30 years. African demographic heritage that is based on population censuses constitutes the most complete source of information, even though it is a threatened heritage, as we will show in later pages. And just what is the situation regarding this heritage in Quebec?

3. Quebec and its demographic heritage: an experience to share

Quebec is often called a paradise for genealogists. Its citizens have easy access to detailed information and this has allowed many persons of French-Canadian origin to find their ancestors over more than 12 generations (Caron, 2002). The potential of the sources preserved in the parishes of Quebec since the start of the French colonization goes way beyond the simple reconstitution of past generations of families. By matching the information taken from parochial registers for marriages, baptisms as well as from tombstones, it was possible to reconnect the demographic customs of the people of Quebec with their past generations (Charbonneau et al., 1987; Vézina et al., 2005).

Nevertheless, the gaps found in registers are now attracting a greater following of researchers and genealogy enthusiasts who try to fill in the deficiencies in information by having recourse to censuses (Marcoux et al., 2003). The quality of the information extracted from registers is not of equal merit, depending on the religious groups present in the population in Quebec. By way of example, the demographic history of the Franco-Catholics in Quebec is much better documented than for other ethno-religious groups. One of the negative effects of such inequality of coverage is that the history of numerous communities and their populations who helped build Quebec was left in the dark. Furthermore, data from registers showed a severe lack of information regarding the activity and work performed by individuals. Often, there was no reference to the level of education, linguistic practices, ethnic origin, or other characteristics essential to have a better understanding of the transformations experienced by the populations at the time.

Canada has not only one of the longest-standing traditions in organizing censuses, but even more, it is one of the rare States to have kept such precious information intact for a long period of time. In keeping with Article 8 of the Constitution Act which established Canada in 1867 (originally called the British North America Act), a general census was required to be taken every ten years beginning in 1871. As of 1951, Canadian censuses operated on a five-year interval. By virtue of Article 17 of the Statistics Act which seeks to protect the confidentiality of all personal information collected, any such information can only be disclosed after 92 years: the 1901 general census was disclosed only in 1993, and the 1911 census in 2003. All the questionnaires filled out by hand by census-taking agents (since 1871 for Canada and 1851 for Quebec) are stored in the national archives of Canada.

The preservation of such documents in paper format presents considerable challenges, such as the need to maintain the physical state of files or the limits to warehousing space. In order to allow for better conservation and to free up warehousing space, the questionnaires were transferred to microfilm in the 1950s (the paper documents were destroyed). The microfilms were in turn replaced some years ago by digitized files (images). Both Canada and Quebec succeeded in safeguarding a vital demographic heritage that contains very detailed information on individuals and families for more than 150 years.

Although the information on microfilm was originally of interest to only a small group of specialists (historians, demographers and other researchers), their widespread use spearheaded by the development of new information technologies has generated a number of initiatives. One of the most important projects now underway aims to give the general public access to all personal information of past censuses on the website of Library and Archives Canada. Using the browser developed by the

Canadian genealogy centre, users can consult written documents relating to each of the 5 million and 7.2 million persons registered in Canada in 1901 and 1911 respectively.

Beyond the circle of amateur and professional genealogists, the scientific communities in Quebec and Canada have joined forces to make great use of the data available. One notable example is the Canadian Century Research Infrastructure (CCRI), a research program headed by professor Chad Gaffield (University of Ottawa) that sets out to study the social, cultural, economic and political changes in Canada by making extensive use of databases from 1911-1951 that were created from personal data in census returns (St-Hilaire, 2009). In 1997, we also embarked on a research project in partnership with our colleague, geographer Marc St-Hilaire, to study the population of the city of Quebec. Using data extracted from the first seven censuses taken in the city (refer to Marcoux et al., 2003; Marcoux and St-Hilaire, 2003; St-Hilaire and Marcoux, 2001 and 2004), we have been able to create a local database with extremely detailed information as well as georeferences to approximately half a million persons who lived in Quebec between 1851 and 1911. All this information is now being processed as part of the research program *Population and social history of the city of Quebec*, and offers us a new vision of the city that celebrated its 400th birthday in 2008: the program looks at the transformations of the city through its modes of residence, death rate, child education and labor, marriage and areas of recruitment for spouses, widowhood and remarriage, fertility rate and family, as well as other factors.

4. Challenges to preserving census data in French-speaking Africa

At the end of the 1960s, conscious of the lack of basic information available on the populations living in the vast majority of the newly independent Sub-Saharan countries in Africa, the United Nations Population Fund (UNFPA) set up the *African Census Analysis Project* (ACAP), which empowered some twenty countries on the African continent to take their first general census (Graft-Johnson, 1988). For reasons of social and political instability, some countries had to wait until the 1980s and 1990s before being able to organize a nationwide data collection operation. Others had already started to gather information on their populations at different periods, as was the case with Mali and Burkina Faso which carried out their fourth census in 2009 and 2006 respectively.

With questionnaires being the only really exhaustive form of data collection employed in African countries, they are also used to record more than 50 characteristics pertaining to individuals and households, with some being compliant with the conventions established by UNESCO and other United Nations agencies: age and sex of responders, national languages and mother tongues; literacy and school attendance rates; matrimonial situations, economic activities, housing conditions and household equipment, etc.

Population censuses in Africa provide the greater part of the basic information used to develop all public policies. The guidelines for economic and social development rely on information generated by these censuses. In short, these vast data collection campaigns form the core elements of all planning exercises and constitute a vital tool for African States, especially for French-speaking countries, many of which plan to organize censuses in the years to come (Annex 1).

Based on the wealth of data now collected, it is safe to say that in the last four decades, Sub-Saharan Africa has managed to extricate itself from the very dire situation of socio-demographic information poverty that it once experienced (Van de Walle, 2006; Marcoux, Zuberi and Bangha, 2005). The problem is that this rapid burst of knowledge of African populations was not accompanied by any real effort to preserve the information collected. Computerized data storage technologies have evolved at

such a rapid pace that, quite often, with no measures having been adopted to transfer the data to new storage media, the data from earlier censuses have now been completely lost — either because the media once used to store such information is now obsolete, or because they have simply disappeared, as highlighted in Gubry and Moriconi-Ébrard (2007, p. 5):

The digital age creates new challenges born from the very rapid evolution of technologies and possible storage conditions. The large magnetic tapes of the Seventies, scheduled to be recycled every ten years or so, were replaced by storage media such as the Bernoulli Boxes which have in turn disappeared, just as their drive readers. This meant that fine-grained diachronic analyses, which would again need to access such information, were no longer possible.

The possibility that data from numerous African censuses could disappear completely is a danger that needs to be addressed, given the financial investments they required. Population censuses are collection operations that need to be carried out by all States, but they are also very costly. It is estimated that the cost of these operations in Sub-Saharan Africa totaled almost one billion dollars for the decade of the 1990s. Funding for a general population census in an African country represents a very significant portion of the limited budgets of most planning ministries and this explains why the international community is often called upon for assistance in such operations. Canada alone has allocated approximately 15 million dollars from its public aid development budgets to finance African census-related activities during the 1990s. It is quite unacceptable that investments of this magnitude are not better protected.

Recent initiatives nevertheless provide reason for greater optimism for the preservation of future census operations. The Integrated Management Information System (IMIS) program set up by the United Nations Population Fund should help ensure that past mistakes are no longer made and that data from censuses taken in the 2000s are better preserved (Zoungrana et al., 2007). With regard to the censuses done in the 1990s, it is important to highlight the initiative of the Integrated Public Use Microdata Series—International (IPUMS) program of the University of Minnesota, in which researchers are allowed to access samples from census databases of some 30 countries worldwide, four of which are in Africa. Albeit an interesting project, this American initiative refers only to census samples (5% to 10%), thus sidestepping one of the primary characteristics of a census, namely the exhaustiveness of the collection process.

A second, noteworthy, American initiative is the African Census Analysis Project (ACAP) of the University of Pennsylvania which helped safeguard the databases of more than 50 censuses from 26 African countries. French-speaking African countries do not however reap much of the benefits of these actions of preservation. This is not new, and is a situation we have already mentioned (Marcoux, 1990; Gervais and Marcoux, 1993).

Nevertheless, we are well aware that in a context where there are deficiencies in the civil status system, censuses represent the most reliable sources of information.

A general population census offers the only way to provide accurate, reliable data as well as other information on said population at every geographical level in African countries. Alternative methods are not practical and there is currently no other source available in Sub-Saharan Africa, nor will there be one in the foreseeable future (Dackam, 2003, p. 96).

If no other options can be devised in the foreseeable future to replace censuses as a sure way to collect detailed information on African populations, **the general censuses of the past constitute indeed, a demographic heritage of prime importance.** Now however, this heritage faces the risk of extinction as,

more often than not, only rare, perfunctory publications exist of the African censuses taken in the 1970s and 1980s: digital databases have completely disappeared, now making it impossible to carry out any new evaluation of past information.

However, certain African countries, following the example of Canada and the United States, have passed legislation making it mandatory to store the handwritten questionnaires of the first censuses in their official archives. This is the case in Mali (Ongoïba, 2007), Burkina Faso (Zoungrana et al., 2007) and the Democratic Republic of the Congo (Lututala et al., 2009). Missions undertaken by ODSEF management have revealed that questionnaires for several censuses have also been preserved in their entirety in Benin, Niger and Senegal. The exchanges we have with statistics and archives managements lead us to believe that this is also the case in several other African countries. These last handwritten, exhaustive vestiges of the lives of the populations in certain countries during the 1970s, 1980s and 1990s are often exposed and deteriorate in the harsh physical conditions of preservation. It is for this reason that more than one hundred researchers, meeting at an international symposium in 2007, signed the *Quebec City Declaration Regarding the Recognition, Protection and Development of African Censuses* (Annex 2).

5. Experiences in Mali and the Congo

One of the signatories of the Declaration was the current director of the statistics institute in Mali (INSTAT), who is also director of that country's archives. It is interesting to note that in 2007, Mali was in the middle of preparations to conduct its 4th national population census. The archives director was quite worried about the idea of having to handle a stock of more than five million, A3-sized documents that represented all the questionnaires for the 2009 census. One of the options under consideration was to destroy the documents from the 1976 and 1987 censuses so as to free up the space needed to accommodate the questionnaires for the 2009 census.

This was the context in which ODSEF and its partners worked to set up a digitization workshop that was inaugurated in January 2010 in the INSTAT facilities in Bamako.² The workshop was created with funding from ODSEF and its partners (technical missions, Mali staff training, etc.) whereas the equipment for the installation was funded by an initial outlay (70 000€) from the Institut de la Francophonie numérique (OIF-University Laval agreement protocol signed in September 2009). The digitizing pilot project for the 1976 census forms in Mali served as a prototype to validate this type of operation in a French-speaking African country. Apart from equipment acquisition, other financial and human resources invested in the pilot project were covered by direct funding from ODSEF and the Mali statistics institute (INSTAT).

During the first three months of operation, the Bamako workshop validated its operations and digitized 20% of the forms used in the 1976 census. Following an ODSEF mission carried out in May 2010, a number of modifications were made to the document preparation process and the production chain. The production rate was increased three-fold, ramped up from a weekly average of 10 000 to 30 000 documents. The work rate was modified, with digitization of the 1976 Mali census completed in January 2011.³

² Information on the installation of the workshop can be found on the ODSEF website (www.odsef.fss.ulaval.ca)

³ Approximately 2 500 000 colour images, in JPEG format at 300 ppp, were produced. The original census forms have recto-verso printing, on paper with dimensions (318 x 450 mm) slightly bigger than an A3 document. A one-centimeter trimming was done in one of the document margins so that high-speed digitizers with automatic sheet-

Given the success of the operation, ODSEF and INSTAT sought to continue their partnership and protect the other two Mali censuses conducted in 1987 and 1998. Whereas the number of forms to be digitized was estimated at one million for the 1976 census, our estimates for the 1987 and 1998 censuses were for 2.5 and 4 million respectively. In order to maximize the chances of success of the operation, we fitted the workshop with a more powerful guillotine and ensured that the computer equipment was operational, or replaced if necessary, at the same time adding a new, high-speed digitizer to start a new production line for images. In order to supervise and support the team in Mali, ODSEF will continue to accompany INSTAT in this project and will carry out staff-training missions to Mali so as to build on the experience acquired from the pilot project for the 1976 census. By mid 2013, the first three population censuses for Mali will have been stored. Copies of the digital files are stored in different locations, including at INSTAT.

It is to be hoped that by the end of 2013, INSTAT will have assimilated the practices of protecting the country's census data into its activities, not only making use of the structure in place (digitizing workshop), but also taking advantage of the equipment and trained personnel to create the requisite conditions to sustain a veritable digital culture in the years to come for the benefit of safeguarding the demographic heritage within the Malian statistics institute and among its partners.

It is worthy of note that with the expertise they have acquired, Mali and the INSTAT team are now the driving force among other statistics institutions in the west African region. Now, teams from Burkina Faso and Niger, with support from UNFPA offices and their respective capitals, have been able to visit Mali on technical missions to tour the Bamako workshop, meet INSTAT partners, familiarize themselves with the different stages of protection and storage, as well as to see the tools used for digitizing census questionnaires.⁴

Just as with Mali, the Democratic Republic of the Congo had also sounded the alarm during a presentation at the Quebec symposium (Lututala et al., 2009). It should be pointed out that the DRC conducted just a single census in its history, in 1984. With the economic and political difficulties facing the country, the data from this census was, for all intents and purposes, left unused (Lututala et al., 2009). In 2009, three institutions from the Democratic Republic of the Congo approached ODSEF to outline their needs and identify the actions to take to ensure protection and recognition of data from the 1984 census. This prompted an ODSEF mission to Kinshasa in September 2010 which led to a series of meetings with the managements of the National Statistics Institute (INS) in the DRC, the University of Kinshasa and the national Congolese Bureau of the UNFPA.

After carrying out a thorough evaluation of the INS archives, ODSEF has stated that the 1984 census documents were safeguarded in their entirety and were in a state that would allow them to be digitized, as was the case in Mali. The problem of document warehousing was also raised by the Congolese authorities. With preparations underway for the next Congolese census in 2014, the 1984 census documents are now at risk, given the limited space that INS has at its disposal for archiving. With a population estimated at 80 million, the storage needs for 15 million questionnaires for the next census

feeders could be used. The same strategy was used for digitizing subsequent Malian censuses and for the 1984 census in the DRC, with both original forms being printed on paper formats larger than A3.

⁴ Cf. two documents: 1) "Bamako mission report, 13 to 22 August 2011," National institute of statistics and demography, Burkina Faso; 2) "Mission report: assessment mission to INSTAT, Mali," Niger institute of statistics (December 2011). The DRC team mission to Bamako should have taken place in March 2012 but has been postponed on account of the events that have gripped Mali.

may incite authorities to remove the 1984 census questionnaires which, based on similar experiences in other countries, could seriously damage the integrity and completeness of these archives.

It is within this framework that ODSEF embarked on a new initiative to set up a second digitizing workshop for census questionnaires, this time in Kinshasa. Despite the numerous constraints present in this African capital, an institutional partnership dynamic and the high motivation shown by the INS staff for this Congolese demographic heritage protection program have combined to create the right conditions for the success of the project. The national bureau of the United Nations Population Fund (UNFPA) covered the cost of outfitting the facilities that will house the digitizing workshop (painting, electricity, air conditioning, office furniture, etc.) whereas a new funding packet received from the Organisation Internationale de la Francophonie was used to cover equipment procurement costs as well as training and project supervision costs by ODSEF (equipment installation, setting up of the production chain, personnel training, etc.). The workshop began operations in May 2012 and after just five months, it has completed the digital storage of the questionnaires for the entire city of Kinshasa and the Lower-Congo province which together account for 15% of the total population of the country. It is estimated that the entire digitization process for the only general population census carried out in the history of the DRC will be completed before the end of 2013.

6. Conclusion

National population censuses are an integral part of the world's demographic heritage. In comparison with other forms of documentary heritage concerning populations, census-taking, as it is conducted in the majority of countries, is singular in its campaign to record the personal details of every individual, irrespective of age, sex, social status, level of education, wealth, class, and so on. This source of information on national populations adorns a very democratic nature, compared to other sources that consider only the life of educated individuals (those leaving written traces), or the elite of a society.

It is a fact that documentary heritage is particularly abundant in some developed countries and especially in Canada. For almost twenty years now, teams from Library and Archives Canada have been working on a monumental project to provide researchers and the general public with tools giving them online access to digitized handwritten documents of Canadian national censuses for a period stretching back more than two centuries. Interest in this type of information is evident for it is one of the most popular sources of information on the website of Library and Archives Canada (Tremblay, 2012).

The countries of Sub-Saharan Africa have for the most part only very recent experience, often dating back to the 1970s and/or the 1980s, in organizing population censuses. Despite the recent nature of African documentary heritage, it is clearly at risk in some countries and urgent action is needed, as was made clear by the *Quebec City Declaration Regarding the Recognition, Protection and Development of African Censuses*, signed in 2007.

Although it has been clearly demonstrated that it is crucial to protect the demographic heritage represented by African population censuses, the very large quantity of documents to process creates sizeable logistics and organizational challenges. Perennial protection of these documents calls for high-speed digitizing equipment, installation of an efficient production chain, pluridisciplinary technical staff with adequate training, and institutional organization coupled with proper supervision to respond in a proactive manner to obstacles that may arise. The experience gained by ODSEF in Mali and in the DRC represents concrete actions that have come in the wake of the Quebec City Declaration. Bolstered by the invaluable collaborative effort with the statistics institutions in these two countries and with the financial

backing from ODSEF partners, it was possible to implement structures to ensure the protection of precious demographical heritage represented by population censuses.

In the coming years, ODSEF will continue to work with its institutional partners at both national and international levels to carry out actions to preserve census data and extend its program to some twenty other French-speaking countries, all this while pursuing four objectives:

1. Ensure the protection of the documentary heritage represented by censuses, a heritage facing a fragile future, but containing insightful social, cultural, economic and demographic information that is unique to African populations.
2. Promote the adoption of a true digital culture in African national institutions in the field of statistics and archives.
3. Promote a regional South-South dynamic through the recognition of digital archival data, and work to support national institutions, the Mali Statistics Institute (INSTAT) and the National Statistics Institute (INS) in the Democratic Republic of the Congo, empowering them to play front-running roles to lead other countries in their respective regions.
4. Conduct collaborative ODSEF actions in partnership with statistics institutions in French-speaking Africa, guiding them in the preparation of future censuses and other surveys, and in so doing, promote the collection of pertinent, quality information, especially as regards national and official languages as well as the linguistic practices of the populations in these countries.

References

- Caron, Caroline-Isabelle. "La narration généalogique en Amérique du Nord francophone. Un moteur de construction identitaire." *Ethnologies comparées*, 4 (2002) "Mémoires des lieux".
- Charbonneau, Hubert, Bertrand Desjardins, André Guillemette, Yves Landry, Jacques Légaré, and François Nault. *Naissance d'une population : les Français établis au Canada au XVII^e siècle*. Paris et Montréal, Institut national d'études démographiques et Presses de l'Université de Montréal, 1987.
- Dackam, Richard. "New Strategies to Improve the Cost-Effectiveness of Census in Africa." In *Counting the People*, Population and Development Strategies Series, 77-97. New York, United Nations Population Fund, 2003.
- Gervais, Raymond, and Richard Marcoux. "Saving Francophone Africa's Statistical Past." *History in Africa* 20 (1993): 385-390.
- Graft-Johnson, K.T. "Les sources de données démographiques en Afrique." In *L'état de la démographie africaine*, 13-28. Liège, International Union for the Scientific Study of Population, 1988.
- Gubry, Françoise, and François Moriconi-Ébrard. "Transformer un gisement en ressources : la valorisation des données de recensement, une mission du CEPED." Paper presented at the international conference "Mémoires et démographie: Regards croisés au Sud et au Nord," Québec, 19-22 June 2007.
- Laboratories for Population Statistics. "Development of a National Strategy for the Improvement of Civil Registration and Vital Statistics." Conference Document 61 PC, Chapel Hill, 1976.
- Lututala, Bernard Mumpasi, Pascal Kapagama Ikando, and Kishimba Ngoy. "État des lieux des données du recensement scientifique de 1984 en République Démocratique du Congo." In *Mémoires et*

- démographie. Regards croisés au Sud et au Nord*, edited by R. Marcoux, 98-101. Québec: Presses de l'Université Laval, 2009.
- Marcoux, Richard, and Marc St-Hilaire. "Régimes démographiques, famille et travail des enfants : y a-t-il une spécificité des nouveaux citoyens d'origine rurale à Québec en 1901?" In *Famille et marché. XVI^e-XX^e siècles*, edited by Christian Dessurault, John A. Dickinson, and Joseph Goy, 323-340. Québec: Septentrion, 2003.
- Marcoux, Richard, Marc St-Hilaire, and Charles Fleury. "Ville et population en changement: transformations urbaines et ajustement familiaux à Québec au XIX^e siècle et au début du XX^e," *L'Ancêtre* 29 (2003): 227-230.
- Marcoux, Richard, Tukufu Zuberi, and Martin Bangha. "Les recensements en Afrique francophone : les enjeux d'un patrimoine menacé." *4^e Colloque francophone sur les sondages*, Québec, 24-27 mai 2005.
- Marcoux, Richard. "Les enquêtes démographiques nationales des années 60 en pays sahéliens francophones." *Bulletin de l'Association canadienne des études africaines* (Winter 1990): 14-21.
- Moon, Jeff. "L'IDD et l'art de former les utilisateurs de données." *Bulletin de l'IDD* 4, no. 1 (2000): 2-6.
- Ongoïba, Aly. "La conservation des archives des recensements au Mali." Paper presented at the international conference "Mémoires et démographie : Regards croisés au Sud et au Nord," Québec, 19-22 June 2007.
- Poirier, Jean and Céline Le Bourdais. 2009. "Bilan et perspectives de l'Initiative canadienne pour les statistiques sociales à la lumière de l'expérience du Centre interuniversitaire québécois de statistiques sociales (CIQSS)." In *Mémoires et démographie. Regards croisés au Sud et au Nord*, edited by R. Marcoux, 148-152. Québec : Presses de l'Université Laval, 2009.
- St-Hilaire, Marc, and Richard Marcoux. "Le ralentissement démographique." In *Atlas historique du Québec. Québec, ville et capitale*, edited by Serge Courville et Robert Garon, 172-180. Sainte-Foy, Presses de l'Université Laval, 2001.
- St-Hilaire, Marc, and Richard Marcoux. 2004. "Quebec City." In *The Oxford Companion to Canadian History*, edited by Gerald Hallowell, 523-524. Don Mills, Ont.: Oxford University Press, 2004.
- St-Hilaire, Marc. 2009. "L'Infrastructure de recherche sur le Canada au 20^e siècle : des outils pour renouveler l'histoire de la population." In *Mémoires et démographie. Regards croisés au Sud et au Nord*, edited by R. Marcoux, 144-147. Québec: Presses de l'Université Laval, 2009.
- Tremblay, Sylvie. "Les recensements numérisés du Canada, une approche nouvelle pour des documents d'archives." Paper presented at the colloquium *La sauvegarde numérique des patrimoines démographiques*, Forum mondial de la langue française, Quebec, July 2012.
- Van De Walle, Étienne. "Introduction" In *African Households. Census and Surveys*, xxi-xxxix. New York, M.E. Sharpe, 2006.
- Vézina, Hélène, Marc Tremblay, Bertrand Desjardins, and Louis Houde. "Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise." *Cahiers québécois de démographie* 34, no. 2 (2005): 235-258.
- Zoungrana, Cécile-Marie, Benjamin Zanou, and Hamissou Kano. "Sécurisation et dissémination des données de recensements au Burkina Faso : pour une meilleure démocratisation de l'information." Paper presented at the international conference "Mémoires et démographie : Regards croisés au Sud et au Nord," Québec, 19-22 June 2007.

Annex 1

Population Census done and planned since 1985 in African francophone countries

Country	1985-1994	1995-2004	2005-2014	<i>Planned after 2010</i>
Algérie	03-1987	<u>06-1998</u>	<u>08-2013</u>	2018
Benin	02-1992	<u>02-2002</u>	11-2012	2012
Burkina Faso	12-1995	12-1996	<u>12-2006</u>	2016
Burundi	09-1990	-	<u>08-2008</u>	
Cameroun	04-1987	-	<u>11-2005</u>	2015
Cap Vert	06-1990	<u>06-2000</u>	<u>(06-2010)</u>	
R. centrafricaine	12-1988	<u>12-2003</u>	(2013)	2013
Tchad	<u>04-1993</u>	-	06-2009	
Comores	09-1991	11-2003	(2013)	2013
Congo	12-1994(1)	<u>07-1996</u>	04-2007	2017
Côte d'Ivoire	03-1988	<u>12-1998</u>	(2012)	2011-2012
RD du Congo	-	-	(2013-2014)*	2012-2013
Djibouti	-	-	<u>12-2009</u>	
Gabon	07-1993(P)	12-2003	(2013)	2013
Guinée	-	<u>12-1996</u>	(2012)	2011-2012
Guinée Bissau	12-1991	-	<u>03-2009</u>	
Madagascar	08-1993	-	(2013)	2011-2012
Mali	04-1987	<u>04-1998</u>	04-2009	2019
Mauritanie	04-1988	11-2000	(2012)	
Maroc	12-1994	<u>09-2004</u>	(2014)	2014
Niger	06-1988	<u>05-2001</u>	(11-2012)	2011-2012
Réunion	03-1990	8 March 1999	<u>01-2006</u>	
Rwanda	08-1991	<u>08-2002</u>	(2012)	2012
Sénégal	05-1988	<u>12-2002</u>	(2012)	2011-2012
Seychelles	08-1987(P)	<u>08-1997</u>	08-2010	
	08-1994	<u>08-2002</u>		
Togo	-	-	11-2010	2011
Tunisie	04-1994	04-2004	(2014)	2014

Source: <http://unstats.un.org/unsd/demographic/sources/census/censusdates.htm#AFRICA>

Annex 2

Quebec City Declaration Regarding the Recognition, Protection and Development of African Censuses



Quebec City Declaration
**Regarding the Recognition, Protection
and Development of African Censuses**



Most African countries have conducted at least two national population censuses over the past 40 years. These censuses have allowed us to enrich our knowledge of African societies. They also provide the most geographically detailed and exhaustive normal data on the population. As such, they constitute the cornerstone of an integrated system of demographic and socio-economic information, and represent a unique legacy in the demographic study of Africa's populations. In contrast to the attention devoted to permanently preserving designated historical documents in archival centres, the comprehensive archiving of census documents and data has been neglected. This data heritage is thereby at risk of destruction, despite its critical value in the production of knowledge on African societies and the monitoring of development policies and programs.

Faced with this situation, various organizations have mobilized to recognize, preserve and develop this legacy. ACAP (the African Census Analysis Project at the University of Pennsylvania) has played a pioneering role in preserving documents. It is also heartening to see the initiatives of CERPOD, IFLMS-International, and CEPED, as well as those developed by the United Nations system, particularly UNFPA and PARIS21, to establish integrated database systems.

Since urgent action is required, we, researchers, educator researchers, and other specialized producers and users of population data, gathered together for the 7es Journées scientifiques du Réseau Démographique de l'Agence universitaire de la Francophonie (AUF), in Québec City from June 19 to 22, 2007, call upon governments, national and international institutions, and the scientific community to ensure that this legacy is protected by taking urgent action to

1. Establish the infrastructure needed for archiving completed questionnaires, cartography, methodological and technical documents, printed reports, computer files, and any other kind of census-related record.
2. Establish a system to facilitate access to data in order to develop research that will promote African expertise, inter/pluridisciplinary approaches, and international comparisons.
3. Develop an inter-institutional partnership based on the use of open protocols and standards that ensure the integration of various resources and their dissemination, taking heritage and copyright considerations into account, and making provision for any required legal protection.

Declaration made in Quebec City on June 22, 2003

PARTICIPANTS AT JOURNÉE SCIENTIFIQUE DE L'ÉCRITURE MIMÉE ET DÉMOGRAPHIE : REGARDS CROISÉS AU SUR ET AU DESSUS

[illegible][illegible][illegible][illegible][illegible]

Source: http://www.demographie.auf.org/IMG/pdf/Quebec_City_Declaration-Census_Africa.pdf

Partnership in Paradise

The Importance of Collaboration for Handling Traditional Cultural Expression Material in the Pacific Islands

Brandon Oswald

Island Culture Archival Support, USA

Abstract

Access to cultural heritage materials in the Pacific Islands has increased along with access to technology. However, the availability of indigenous Traditional Cultural Expression (TCE) information causes concerns regarding the protection of cultural sensitivities. Digitization plays a major role in the accessibility of this material, as there is presently no control over its dissemination. This paper will first examine the structure and meaning of TCE. Collaboration with indigenous people is vital because solutions may reside in the development of mutually satisfying pathways for the future management of such valuable material. Since many of the cultural heritage organizations in the Pacific Islands constantly seek to develop new frameworks for caring and making accessible TCEs, this paper will then review possible strategies and guidelines that archivists in the Pacific Islands can utilize to better safeguard access and control.

Author

Brandon Oswald is currently the Founder and Executive Director of the non-profit organization, Island Culture Archival Support (ICAS) that is dedicated to providing voluntary archival assistance to cultural heritage organizations in the Pacific Islands. He began his archival career in the Digital Library Unit at the Scripps Institution of Oceanography Library before turning to freelance work. Brandon has a Master's in Archives and Records Management from the University of Dundee. He is an active member of the Pacific Regional Branch International Council on Archives (PARBICA).

*“E lauhoe mai na wa’a; i ke ka, i ka hoe; i ka hoe, i ke ka; pae aku i ka ‘aina”
“Everybody paddle the canoes together; bail and paddle, paddle and bail,
and the shore is reached.”¹*

--Hawaiian Proverb

1. Introduction

Although traditional cultural expressions (TCEs) typically do not have precise guidelines, they are commonly accepted as works that reflect and identify a community's history, cultural and social identity. They typically imbue values that are handed down from one generation to another, either orally or by imitation. They are often made by unknown authors, or by individuals within a community who are recognized to have the right, responsibility or permission to create them. TCEs are considered by native peoples to belong to the community in which they originate, and are integral to self-determination.² This

¹ From Huna Website: <http://www.huna.org/html/inspire/reply13.html>.

² Janice T. Pilch. “Issue Brief: Traditional Cultural Expression,” 1 September 2009, <http://www.librarycopyrightalliance.org/bm~doc/issuebrieftce.pdf> (6 June 2012).



Figure 1. Map of the South Pacific Islands.

has never been more evident than in the Pacific Islands, where TCEs are often considered sacred and not meant to be shared outside the community.

The Pacific Islands possess diverse and vibrant cultures. Since their discovery in the late eighteenth century, they have attracted a myriad of people such as adventurers, evangelists, artists, and researchers—who described the Pacific Islands’ as a kind of a natural society unburdened by civilization’s problems and used it as a model and counter-point to advanced civilization. These people have captured TCEs

creating an innumerable amount of material in the form of correspondences, field notes, film, photographs, audio recordings, journals, paintings, drawings, genealogies, and diaries. Since the advent of the Digital Age in the latter half of the twentieth century, technology has made it easier for these people to repatriate them to cultural heritage organizations such as archives, libraries and museums, where accessibility and use of TCE material is done without consulting with indigenous groups, or acknowledging customary laws and intellectual property rights. In fact, when non-indigenous people own most of the material that documents indigenous people's lives and traditions legally certain tensions arise.³ These tensions are expressed on matters relating to access and control.

The strengthening of communication and the building of new relationships between cultural heritage organizations and indigenous peoples and traditional communities depends on how actively both parties will develop, manage, and maintain strategies concerning the safeguarding and access to TCE material. This relationship is vital, as solutions may reside in the development of mutually satisfying pathways for the future management of such valuable material. Tradition-bearers can provide invaluable contextual information and personal narratives regarding collections about them. Indeed, indigenous peoples and traditional communities do have a legitimate interest in being part of the decision-making process regarding TCE material. Today, this voice is essential, as it can explain alternative meanings embedded within them. They can help outline the access conditions that respect the indigenous or traditional community from which those materials derive, as well as those of other users who are keen to learn and understand different cultures and cultural practices from them.⁴

Indeed, in the past several decades there has been a publicized absence of trust between indigenous and traditional communities and the cultural institutions, as they feel that they have not been recognized as rights holders or acknowledged as having legitimate relationships with the material within the collections. Additionally, professional volunteers come from countries outside the Pacific Islands region to work in a variety of ways at cultural heritage organizations. However, these volunteers may not understand all the nuances of working with TCE material that can be a source of conflict. As this paper will show, it will only benefit both indigenous and traditional community and the cultural heritage organization to work together beyond the risks and troublesomeness in order to understand how to best protect and make accessible their rich cultural heritage.

2. The Issus with Traditional Cultural Expression

TCEs embody know-how and skills, and transmit core values and beliefs. They are integral to the cultural and social identities of indigenous and traditional communities. As cultural and economic assets, their protection is linked to the promotion of creativity, enhanced cultural diversity and the preservation of cultural heritage. TCEs include music, art, designs, names, signs and symbols, performances, architectural forms, handicrafts and narratives. In many countries where knowledge is transmitted in oral form, particularly in the Pacific Island countries, traditional knowledge and expression of indigenous cultures are a living and evolving tradition. Such knowledge and expressions are socially based and communally

³ Jane Anderson, "Indigenous Knowledge, Intellectual Property, Libraries and Archives: Crisis of Access, Control and Future Utility," in *Australian Indigenous Knowledge and Libraries*, ed. Martin Nakata and Marcia Langton (Canberra: Australian Academic & Research Libraries, 2005), 90.

⁴ Molly Torsen and Jane Anderson, *Intellectual Property and the Safeguarding of Traditional Cultures: Legal Issues and Practical Options for Museums, Libraries and Archives* (World Intellectual Property Organization, 2010), 12.

owned because they have been developed for the benefit of the group as a whole.⁵ Often, it is a group that is the custodian of a particular item of heritage, and thus, the consent to authorize others to use indigenous cultural knowledge must be given by the group, through specific decision-making procedures which differ depending on the nature of the particular item.

As cultural heritage organizations in the Pacific Islands region are drawn into the digital age, managing access and use of collections inevitably involves intellectual property law, policy and practice. The very nature of traditional cultural expressions means that they occupy an ambiguous intellectual property status. Unique intellectual property issues are being raised regarding collections of TCE because of the certain qualities that make them different from other collections. Traditional knowledge and indigenous cultural expression is vital for the survival of the cultural identity of Pacific Island nations and there is a real need to protect local cultures to ensure that there is something to pass on for the future.

Contemporary negotiations over the rights of indigenous peoples' and traditional communities' and interests in their TCE raise a number of challenges for cultural heritage organizations. Often these challenges arise from the complex social, historical, cultural, legal and political conditions informing the collections of these organizations, and can manifest in a variety of ways. Firstly, researchers often collect TCEs without obtaining consent from the source communities, as indigenous people typically do not understand as to how the information will be used. These collections may contain secret or confidential material that may be subject to restricted use under customary laws and practices. Secondly, the legal status of TCEs under intellectual property law is often unclear. A good example of this is when a cultural heritage organization makes accessible material that they believe exists in public domain and considered free to be used by anyone, but, in fact, may be subject to certain restrictions according to the source community. Finally, indigenous peoples and traditional communities wish to be more directly involved in the decision-making processes related to the management of their cultural heritage. This involvement will help them reconnect with those elements of their cultural heritage, allow greater access for children and the community as a whole, and revive the knowledge systems associated with these elements.⁶ On the other hand, indigenous peoples may wish to regain full possession of their TCEs and bring them back to the community.

3. Secrecy and Sacred Material

The term "secret," or "sacred" material is best defined as objects having a particular value for members of the originating community.⁷ There are both published and archival materials in many cultural heritage organizations of the Pacific Islands region that hold secret and sacred information which should not be made generally available. These objects command respect and therefore require special care or the observation of prohibitions, and thus the treatment of these objects by cultural heritage organizations and electronic media requires prior consultation. Often, Pacific Islanders have given secret information to

⁵ Terri Janke, "Pacific Indigenous People Unite to Protect Cultures: Report on the Symposium on the Protection of Traditional Knowledge and Expressions of Indigenous Cultures in the Pacific Islands," (Symposium on the Protection of Traditional Knowledge and Expressions of Indigenous Cultures in the Pacific Conference, Noumea, New Caledonia, 15-19 Feb. 1999), 3.

⁶ Torsen and Anderson, "Intellectual Property and the Safeguarding of Traditional Cultures," 11.

⁷ Australian Institute of Aboriginal and Torres Strait Islander Institute, "Aboriginal and Torres Strait Islander Library and Information Resources Networks Protocols," <http://www1.aiatsis.gov.au/atsilirn/protocols.atsilirn.asn.au/ATSILIRNprotocols.pdf> (6 June 2012).

respected researchers, not realizing that the information would be published and made available to the general public. Secrecy continues to be a double-edged sword that defends powerful knowledge but also imperils the reliable transmission of cultural information. Often, private knowledge and tradition are held by only a handful of anointed experts- those who are at the top of the social ranking, and who are trusted to pass their knowledge to the appropriate person(s). However, with their unexpected deaths, this knowledge and tradition can be lost. Ironically, in a surprising number of cases, information “stolen” by missionaries and anthropologists has saved indigenous communities from tragic cultural losses.⁸

In the Pacific Islands, knowledge was traditionally held more valuable than material possession. While material goods held a transitory value and were distributed to everyone, information was non-distributed wealth, jealously guarded and passed down from father to son.⁹ Today, this is still true, and is especially practiced in governments throughout the region. For example, knowledge is strictly controlled by government hierarchies, as information flows from the bottom up, and orders for the information from the top down. Generally, senior staff must ask for information before it is offered. On the other hand, junior staff members do not volunteer or ask for information from superiors without permission to do so. Thus, the ministers control information within their sectors and lateral transfer of information is prohibited without special accord.

Indigenous peoples insist that documents contain sacred knowledge so authentic and powerful that access to them should be carefully controlled.¹⁰ Often, the distrust that many indigenous and traditional peoples feel towards cultural heritage organizations will deepen, as the dissemination of sacred material (deliberately or innocuously) happens. However, cultural heritage organizations making available online images of material in their collections without proper consultation with source communities run the risk of inadvertently displaying to the widest possible audience secret and sacred material that should not be publicly seen at all. On the other hand, it is clear that cultural heritage organizations perceive databases as a form of repatriation and not as forms of insult, offense and threat to traditional cultural values.

As far as it is known, there exists no exhaustive list of “sacred” or “culturally sensitive” objects for any given community throughout the Pacific. Secrecy is inherently threatening the democratic process and to the public good except in a sharply circumscribed range of situations, for instance, when an Archive abides by a donor’s request that material be closed to researchers for a stated period of time. Additionally, the lack of such exhaustive lists may also indicate disagreements within the communities themselves, as community spokespersons should have the right to check everything that is to be shown and written concerning them, for example, historical details that they judge to be objectionable.

4. Who Then Owns Culture?

As indigenous peoples and traditional communities of the Pacific Islands desire access to existing cultural material so that it can be reinterpreted and given new meaning, the question as to who should be entitled to make decisions concerning such material is a contentious debate. Should the researcher, community or the cultural heritage organization be the rightful owner of the cultural material? This issue creates a bitter and intense atmosphere in relations with indigenous communities as wells as with the cultural heritage

⁸ Michael F. Brown, *Who Owns Native Culture* (Cambridge: Harvard University Press, 2003), 31.

⁹ “Information Constraints in the Pacific Islands,” (Ministerial Conference on Environment and Development in Asia and the Pacific 2000, Kitakyushu, Japan, 31 August – 5 September, 2000), 2.

¹⁰ Michael F. Brown, “Can Culture Be Copyrighted?” *Current Anthropology* 39, no. 2 (April 1998): 193-222, p. 201.

organization that holds it. It is difficult on the organization, in terms of time and labor, but may also be considered quite unfair to the community that wants to use material that represents them and their culture.

It is clear that the concept of ownership (either by individuals, families or communities) in the Pacific Islands region of songs, dances and other forms of traditional knowledge and custom has been well established for a long period of time. In the past, knowledge of styles of singing and dancing, of sculpting slit-gongs or weaving mats, of myths of origin told in local languages, together with the associated rights of performance was a commodity exchanged between local groups. As Dr. Jacob Simet, Executive Director of the Papua New Guinea National Cultural Commission stated:¹¹

We have had songs, traditional knowledge and so on for hundreds of years. There was no doubt as to who originally owned them- they were originally owned by one person who passed them on to his or her clan. There were clear customary laws regarding the right to use the songs and the knowledge. There was no problem in the past.

Today, confusion often arises when people living non-traditional lives claim to be indigenous and try to assert the special protection that might have been appropriate to the indigenous status of their ancestors. This particularly is an issue in Hawaii, as knowledge created in ancient times, belonging to an indigenous group as a whole, is held today by individuals who are fully assimilated into the American culture. Property rights, for example, belong to the people who own the property, and not to the property itself. Thus, intellectual property rights of an indigenous group are not written down in copyrights, patents, or trademarks; but need to be treated as though they were.¹²

Perhaps, the best answer to the question as to who truly owns TCE material exists in intellectual property rights, as any given item in a collection has an intellectual property status. TCE may or may not benefit from intellectual property, thus managing access to and use of material inevitably implicates intellectual property law, policy and practice. Be that as it may, cultural heritage organizations sit at the junction between tradition-bearers and the public. In their daily activities they have a unique opportunity to allow the public to access cultural heritage, and at the same time protect the TCEs and preserve the rights and interests of their bearers. Therefore, today, some cultural heritage organizations and communities in the Pacific Islands region have developed intellectual property related policies and practices concerning the access, ownership, control of content, and safeguarding of cultural heritage. The Vanuatu Cultural Center is exemplary in this practice, as precautions in the manner of forms must be completed and approved for access to certain kinds of TCE material.

5. Reasons for the Protection of Traditional Cultural Expression

Protection in the intellectual property sense is distinguishable from the “safeguarding” or “preservation” of cultural heritage and expressions, but can complement them.¹³ Safeguarding in the context of cultural heritage typically refers to identification, documentation, revitalization and promotion of real and virtual

¹¹ Miranda Forsyth, “Intellectual Property Laws in the South Pacific: Friend of Foe?” *Journal of South Pacific Law* 7 (2003): 2.

¹² Kenneth R. Conklin, “Indigenous Intellectual Property Rights- General Theory, and Why It Does Not Apply in Hawaii,” *Angelfire*, 2011, <http://www.angelfire.com/hi2/hawaiiansovereignty/indigenousintellproprts.html> (6 June 2012).

¹³ Brown, “Can Culture Be Copyrighted?” p. 201.

cultural heritage material in order to ensure its maintenance and essence. However, archiving, documenting, and recording are typically means in which protection of material is achieved. This will mean acknowledging and giving effect to the broader range of collective and individual rights that are linked to TCE. Additionally, there will be a need to build a capacity to support traditional creativity and the social structures that sustain and express them.

The question, ‘what do indigenous peoples seek to protect?’ is not easy to answer, as reasons for seeking protection will be specific to the particular information or knowledge for which the protection is being sought. For example, European connections with haka, a Maori traditional dance, are longstanding in New Zealand. Maori community leaders began to take a public role in supporting haka performances and to insist on correct usages in accordance with traditional custom. In fact, customary standards continue to exercise some control over limited roles for women in haka performance, over performances by foreigners, and over commercial uses. However, in recent years, certain perceived misuses of haka in advertising overseas have become a major source of contention. In New Zealand, the advertisements were widely thought to be culturally insensitive to Maori as well as to New Zealanders. And although Maori groups tried to intervene and ask the producers to change the advertisements, their supplications were mostly ignored.

When indigenous knowledge is removed from an indigenous community, the community loses control over the way in which it is represented and used. Nevertheless, a more general concern is not just with an individual sign, symbol, story, song, or artwork but also with the knowledge and cultural, spiritual and other significance that the material conveys and carries with it. In Australia and New Zealand indigenous people have voiced four major reasons for protecting of TCE material that would also be applicable to the Pacific Islands region in general:¹⁴

- Safeguarding traditions that depend on secrecy, or cultural privacy. Access and use is available only by qualified people or with appropriate permission.
- Maintaining cultural associations in public eye, or cultural publicity. Use is sought to include acknowledgement of cultural associations and by qualified people or with appropriate permission.
- Commercial exploitation of resource or its application treated as commodity, or cultural property. Use is only for agreed purpose and usually on terms of compensation by qualified person or with appropriate permission.
- Maintaining guardianship of traditional culture as a whole. Any use of material must respect and be consistent with guardianship and be by a qualified person or with appropriate permission.

6. Recommendation Guidelines for Repatriated Material

Records relating to excavated or destroyed archaeological sites, societies that have fundamentally changed, life histories of individuals and communities, languages no longer spoken, cultural material that no longer exists, texts of oral renditions of historical events and statements of worldview, all constitute a

¹⁴ Christopher Antons, ed. *Traditional Knowledge, Traditional Cultural Expressions and Intellectual Property Law in the Asia-Pacific Region* (Austin: Wolters Kluwer, 2009), 278.

tenuous link to knowledge that is otherwise irretrievable.¹⁵ Saving TCE materials involve ethical considerations aside from the responsibility for preserving the information they contain. Each cultural heritage organization must consider both access and confidentiality issues, as well as plan for legitimate access before transferring papers to cultural heritage organizations, or even posting digitized copies on the Internet. The scholar who gathered and created the materials must review them to note any sensitive items. Additionally, it is important to remember that situations do change; what is sensitive at a given time may not be later and vice versa. The needs of future researchers and members of the communities are important, but specific decisions will be based on the type of record and the original research purpose.

Cultural heritage organizations often hold original records that were created by, about or with the recommendations of particular Pacific Islands' communities. Some records may have been taken from the control of the community. A practical strategy that allows communities to have more rights to material when historically they have had none is very important. Communities may place tremendous importance on particular records and request copies for use and retention within the community. Thus, to address this issue, cultural heritage organizations should:

- Respond sympathetically and cooperatively to any request from Pacific Islander communities for copies of records of specific relevance to the community for its use and retention.
- Agree to the repatriation of original records to Pacific Islander communities when it can be established that the records have been taken from the control of the community.
- Seek permission to hold copies of repatriated records but refrain from copying such records should permission be denied.

Cultural heritage organizations need to understand the importance of preserving and making accessible the various complementary forms of TEC materials. This understanding will require active participation with donors and the indigenous communities, as well as the consideration of rights of the donor of the material and their wishes or the wishes of their heirs. Ready access to the records can be vital to the indigenous community's knowledge of itself. It certainly is not uncommon for records to be held in distant, often unapproachable institutions where they are effectively alienated from the people to whom they are most relevant. This issue will allow an opportunity for collaboration between indigenous communities and the cultural heritage organization to consider the appropriate location for records with possible cooperation in the development of community keeping places like an archive, library or museum.

7. Recommended Digitization Guidelines

Digitization provides opportunities to improve Pacific Islands indigenous peoples' access to historical and contemporary cultural indigenous knowledge materials that are currently dispersed in institutional collections across the region, and as well as across the world. However, easier access provided by digital technologies also increases the risk of breaching indigenous cultural protocols for the management of materials. Thus, developments in both the digital context and the indigenous information context indicate the need for a coordinated planning approach to deal with the issues. Guiding principles should include:

¹⁵ Nancy J. Parezo, "Preserving Anthropology's Heritage: CoPAR, Anthropological Records, and the Archival Community," *The American Archivist* 62, no. 2 (Fall 1999): 271-306, p. 290.

- Recognize indigenous communities as equal partners in future digital collaborations.
- Ensure cultural integrity.
- Uphold cultural intellectual property rights of communities.
- Interpret, analyse, and synthesize information for general users.
- Ensure that individuals should only access sacred, genealogical information after they have consulted with the relevant indigenous members.
- Ascertain that sacred information is not used in a manner contrary to indigenous communities' values, or for commercial purposes.
- Ensure that information cannot be changed or altered unless after consultation with relevant indigenous people or communities.

Digital databases with potentially vast public access must use sensitivity in the development of their guidelines, determine which material will be made available and establish levels of access that can be implemented.¹⁶ These areas require much attention and discussion prior to the establishment of the digital database. However, as anticipated, the challenge for the development of the digital database remains the extent to which they accommodate the desires and expectations of tradition-bearers about access and control within their governing guidelines. Nevertheless, as technology constantly changes, the preservation of digital databases will also invariably change. Thus, it will be advantageous if the guidelines are flexible and can change with this technology so that it continues to fit the needs of the indigenous community.

Digitization can be an expensive and technically complex process, and it is not uncommon for cultural heritage organizations to fall behind on projects. Moreover, ancillary costs associated with establishing and maintaining an online presence to facilitate public access to images can become a major burden on an organization. Therefore, one bold strategy that will combat against this issue is through the development and maintenance of a distributed contextual information framework. This framework would be designed to link together sources of knowledge that may be of value to a future society.¹⁷ The mapping of these entities (such as, people, organizations, concepts, ideas, places, natural phenomena, events, and cultural artefacts) creates a network of nodes and arcs that mimic actuality. In other words, this creation and caretaking of digital cultural heritage resources utilizes the Internet that enables the linking and sharing of information. An excellent example of this is the Oceania Digital Library, focused on the culture and heritage of the indigenous peoples of the Oceania region.¹⁸ The use of this open-source technology will help cut costs, and can even include individuals with limited technological understanding.

Although these kinds of open-source databases are becoming more popular, very few are being created in the Pacific Islands region. Typically, these digital projects are a collaboration of universities that are inspired to digitize, preserve and provide searchable access to a range of cultural and heritage resources from research collections of partner institutions. Unfortunately, the extent of participation that

¹⁶ Torsen and Anderson, "Intellectual Property and the Safeguarding of Traditional Cultures," p. 73.

¹⁷ Gavin McCarthy, "Finding a Future for Digital Cultural Heritage Resources Using Contextual Information Frameworks," in *Theorizing Digital Cultural Heritage: A Critical Discourse*, ed. Fiona Cameron and Sarah Kenderdine (Cambridge: MIT Press, 2007), 251.

¹⁸ See Website, <http://www.oceania-digital-library.org/>.

indigenous communities are making towards these projects is uncertain. Nevertheless, there are a number of benefits in using an open-source database approach, which include:¹⁹

- Making knowledge transfer easier and encouraging the sharing of experiences within and between organizations.
- Enabling existing knowledge sources to reside within a structured and visible system.
- Increasing visibility of knowledge sources that will promote their preservation and value.
- Introducing nonintrusive techniques that complement existing business practice.
- Supporting the decision-making process by making available a wide range of information giving a view of “the bigger picture.”
- Improving transparency within the cultural heritage community and for external observers (transparency is fundamental to building trust).
- Referring to sensitive information within the system without it being reproduced.
- Improving discovery, accessibility, and comprehensibility of resources.

As we live in a world of perpetual change, there is a very real threat that the accumulated knowledge that guides the digital cultural heritage resources today will be lost to the next generation. Nevertheless, on a positive note, there is no one body that has overall control or the responsibility for overseeing the placement of information on the World Wide Web. It is envisioned that contextual information framework would adopt this strategy, and thus evolve the robustness that comes with dispersed but shared responsibility. Therefore, indigenous communities could establish themselves with cultural heritage organizations as the principal nodes in the network, providing major resources for their communities and a management function in the area of standards.

8. Managing Intellectual Property Guidelines

Indigenous cultural and intellectual property management must be articulated from the start of any digitization project. The main legislative deficiency affecting digitized projects is the temporary and individualistic protection that copyright offers to creators of TCE material. Cultural heritage organizations are in a unique position of being both copyright owners and copyright users. As a result, the primary rights of owners must be recognized. Thus, cultural heritage organizations need to:

- Become aware of the issues surrounding cultural documentation and the need for cultural awareness training.
- Create ways, including the recognition of ethical rights, to protect indigenous peoples and traditional communities.
- Develop proper professional recognition of the primary cultural and intellectual property rights of indigenous people and traditional communities.
- Share information on initiatives involving cultural documentation.

¹⁹ McCarthy, “Finding a Future,” p. 254.

Personal information or culturally sensitive material should not be widely circulated. Typically, sensitive materials include those relating to family photographs, family trees or any recorded historical information and past histories. This type of material should be treated as sensitive and access to this information should be carefully monitored. In fact, appropriate management practices will depend on both the materials and the communities served by the cultural heritage organization. In implementing the processes through which such materials are managed, cultural heritage organizations should:²⁰

- Consult in the identification of such materials and the development of suitable management practices with the most appropriate representatives of the particular indigenous traditional community involved.
- Facilitate the process of consultation and implementation by developing effective mechanisms including liaison with reference groups at local and national levels.
- Seek actively to identify the existence of secret or sacred and sensitive materials by retrospectively surveying holdings and by monitoring current materials.
- Provide suitable storage and viewing facilities with limited access as may be required.
- Ensure that any conditions on access are understood by staff and users and are fully implemented.
- Support the establishment of a national database for the identification of publications with secret or sacred content and of suitable management practices.

Appropriate handling will typically mean making users aware of the contents before they open them. This might involve the need to create labels or notes in the database indicating that the contents are, for example, “For initiated males only,” or any kind of note that states who is allowed to look at the documents.

It is inevitable in today’s current environment that TCE will be disseminated on the Internet in ways that will agree with the wishes of its custodians and in ways that will not. The Internet offers tremendous potential for people to learn more about indigenous culture. However, if respect is to be maintained, guidelines are needed to shape proper conduct in relation to TCE. Cultural heritage organizations must ensure that TCE material is not released unless it is in agreement with best practice guidelines. The development and implementation of guidelines for dealing with TCE materials is becoming an important means of ensuring that the rights of indigenous peoples are recognized. Guidelines will encourage ethical conduct and promote interaction based on good faith. Although protocols are not legally binding, they will establish practices that can, over time, come to be regarded as industry standards.

9. Awareness and Education for Professionals in Cultural Heritage Organizations

The cultural heritage organization’s role as an educational source will also benefit by addressing the concerns of scholars and academics regarding access. This, in turn, can help promote the vitality of indigenous peoples and cultures. Increasing awareness of indigenous communities and their TCE material will only contribute to a greater understanding between the indigenous communities and the cultural heritage organization. In pursuit of awareness, cultural heritage organizations should:²¹

²⁰ Australian Institute of Aboriginal and Torres Strait Islander Institute, “Aboriginal and Torres Strait Islander Library and Information Resources Networks Protocols,” accessed 6 June 2012, <http://www1.aiatsis.gov.au/atsilirn/protocols.atsilirn.asn.au/ATSILIRNprotocols.pdf>.

²¹ Ibid.

- Be proactive in the role of educator, promoting awareness of indigenous peoples, communities, cultures and issues among non-indigenous people.
- Be proactive in acquiring materials produced by indigenous peoples and communities.
- Highlight indigenous content and perspectives through such means as oral history and record copying projects.
- Promote awareness and the use of indigenous related holdings, by such means as targeted guides, finding aids, tours and exhibitions.
- Promote the inclusion of indigenous community peoples to governing and advisory boards, councils, and committees.

Cultural heritage organizations will also greatly benefit by ensuring that their staff are appropriately prepared to deal with TCE material. All graduates of education and training programs for cultural heritage organizations should have gained an appreciation of Pacific Islander history and culture and of the issues relating to the documentation that they will handle in their future careers. Therefore, archives, libraries and museums of the Pacific Islands region should strive to:

- Provide indigenous cultural awareness training for every staff member, especially those who handle TCE material.
- Provide appropriate models for professional practice in acquisition, processing, and collection management on matters concerning indigenous peoples and traditional communities.
- Ensure that education and training programs involve indigenous peoples in both design and delivery.
- Support indigenous students in archive, library and museum education and training through such means as positive encouragement, mentoring and study leave.

Additionally, cultural heritage organizations can help indigenous people promote awareness and make available their histories and perspectives by a number of ways. First, is to mount displays that incorporate, or that are undertaken by indigenous people. A second way is for an archives, library, or museum to host guest speakers or hold community nights where indigenous people tell stories or present cultural performances. Finally, as publishers, cultural heritage organizations can help indigenous peoples to make available their histories and perspectives. Oral history, for example, can result in the publication of tapes, videos or books.

Finally, proper awareness and educational practices will ensure the involvement and participation of indigenous communities in governance, management and operation. It is now widely accepted that all cultural heritage organizations should adopt a client focus, and should respect and solicit the views of their clients. Indeed, these organizations have a certain responsibility to the communities and nations they serve which make it imperative that the communities' interests are reflected in both their governance and management to ensure that all policies and practices serve the interests of the communities without discrimination.

10. Conclusion

Although indigenous people operate many cultural heritage organizations of the Pacific Islands region, there remains a need for archivists, librarians and museum curators to understand the different cultures, which their organizations encounter and represent, as well as be sensitive to their continual development to suit changing needs. Thus, archival development is a necessity if documentary records are to take their place alongside oral evidence in the fabric of island identity.²²

Some of the materials placed online are of sensitive or secret nature that should require certain restrictions on access. Indeed, archivists, librarians and museum curators are caught in a dilemma, as calls for organizations to honor indigenous rules of secrecy have affected everyday practices in their repositories. In fact, some archivists confess privately that they have relocated sensitive records to storage areas that casual visitors are unlikely to find.²³ Nevertheless, it should be the responsibility of each cultural heritage organization that holds TCE material to seek guidance on access, handling and storage. For example, the creation of simple permissions templates can be developed that details those that were consulted, what access was granted, and who should be contacted for further information.

Unfortunately, indigenous and traditional communities of the Pacific Islands are generally among the poorest and most disadvantaged in the world and concern over properly dealing with TCE is not simply an ethical issue.²⁴ Therefore, it is recommended that cultural heritage organizations of the Pacific Islands region need:²⁵

- Assistance in setting up clear institutional infrastructures and systems for collecting, recording, storing and interpreting cultural heritage material;
- Assistance in the training of staff in more up-to-date systems of collecting, storing and recording of cultural heritage material;
- Assistance in the training of staff on intellectual property issues, generally and specifically in relation to museums and archives;
- Assistance in developing and formulating good practices that will guide staff in collecting institutions about how to deal with intellectual property issues; and,
- Availability of funding for inventory projects.

Perhaps, one of the biggest themes found throughout this research was the importance of cultural heritage organizations' interaction with indigenous peoples and traditional communities prior to placing digitized material in an online database. When it comes to digitization and the Internet, guidelines that deal with intellectual property and technology issues are needed for the sustainable management of indigenous materials. These guidelines will help cultural heritage organizations develop coordinated policy approaches. Undoubtedly, indigenous and traditional communities want to be more involved in the process. And as they have legitimate opinions about how TCE collections should be managed, they are beginning to help set guidelines, codes of conduct and protocols. They are even helping to set standards and safeguards for how research should be conducted, what intellectual property rights will remain with the community and where permissions for use in the future will be required. For many cultural institutions

²² Evelyn Wareham, "From Explorers to Evangelists: Archivists, Recordkeeping, and Remembering in the Pacific Islands," *Archival Science* 2, no. 3-4 (2002): 203.

²³ Brown, *Who Owns Native Culture?* p. 31.

²⁴ Torsen and Anderson, "Intellectual Property and the Safeguarding of Traditional Cultures," p. 14.

²⁵ Australian Institute of Aboriginal and Torres Strait Islander Institute, *ibid*.

the problems arise because there is insufficient information about the source community. This hinders negotiations and consultations about the access and use of TCE material by third parties.

“Heritage is not lost and found, stolen and reclaimed. It is a mode of cultural production in the present that has recourse to the past.”²⁶ By definition, heritage looks backward. It is something received from one’s forebears.²⁷ Pacific Islands’ indigenous cultures continue to flourish today. Unfortunately, due to attempts at integration and assimilation of island peoples into Western culture, much traditional knowledge has been lost or diluted. However, material produced by traditional communities and captured by others will offer a rich source for Pacific Islanders to re-connect with their traditional cultural roots. In fact, they provide evidence of stories, dances, languages, and other traditional knowledge that may be at risk of disappearing. This risk becomes much more mitigated, as cultural heritage organizations work together with traditional communities and their TCE material.

²⁶ B. Kirshenblatt-Gimblett, quoted in Brigitte Derlon and Marie Mauzé, “Sacred or Sensitive Objects,” p. 1, accessed 6 June 2012, http://www.necep.net/papers/OS_Derlon-Mauze.pdf.

²⁷ Brown, *Who Owns Native Culture?* p. 183.

Ensuring That it Won't Happen Again

Financial Records and Their Discontents

Safeguarding the Records of our Financial Systems

Victoria L. Lemieux

*School of Library, Archival and Information Studies, The University of British Columbia, Canada,
vlemieux@mail.ub.ca*

Abstract

Financial records are documentary heritage as important as any other form of documentary heritage. Thus the future understanding of our times will very much depend on the preservation of our financial records. The world of finance is increasingly a digital one. In this new world, wealth is generated in digital “financial memory systems.” This paper explores four examples of this debasement of financial memory systems along the supply chain for the now infamous U.S. private label residential mortgage-backed securities: the failings of the mortgage origination process; the failings of MERS—the U.S. Mortgage Electronic Registration System that was created to facilitate the securitization of millions of residential (many sub-prime) mortgages; the process of rehypothecation—when broker-dealers use client money for their own—and associated recordkeeping failures; and the fate of the records of Lehman Brothers after the bank collapsed and the difficulties that this has created for those trying to recover their money. The final part of the paper will shift focus to the actions that need to be taken in order to build new and strengthened financial memory systems—the next generation cyberinfrastructure—for the world’s financial markets.

Author

Dr. Victoria Lemieux is an Assistant Professor at the University of British Columbia (UBC), School of Library, Archival and Information Studies (SLAIS). Her interest in financial records and their relationship to risk stems from her 1999-2001 doctoral research on the information-related causes of the Jamaican Banking Crisis (University College London 2002). Following completion of her doctoral research, Dr. Lemieux joined Credit Suisse, a global investment bank where she worked on various aspects of risk management. In 2008, upon joining UBC, she established the Centre for the Investigation of Financial Electronic Records (CIFER Research). She also serves as a Canadian representative to the International Standards Organization’s Technical Committee on Financial Services (TC68) and is the Acting Director of UBC’s Media and Graphics Interdisciplinary Centre (MAGIC).

1. Introduction

Financial records form an important part of our global documentary heritage. We remember certain historical periods for religious reform, others for their art, and still others for slavery, wars, genocides, etc. This historical period, however, we will remember for the impact of financial decisions on global financial stability. Thus the future understanding of our times will very much depend on the preservation of our financial records.

The world of finance is increasingly a digital one. In this new world, wealth is generated in the 1s and 0s that fly around the globe through an almost unimaginable array of financial systems. The Peruvian economist Hernando de Soto has written that “The reason credit and capital have contracted for the past five years in the U.S. and Europe is that the knowledge required to identify and join parts profitably has

been unwittingly destroyed.” As de Soto explains, in 2008, we began to learn that our financial “memory systems had stopped telling the truth.”¹ This, de Soto claims, is owing to the fact that, for the past 15 years, the records of western capitalism have been debased, leaving institutions and governments without the facts to spot what needs to be fixed or to see where risks existed.²

This paper briefly explores four examples of the debasement of financial memory systems along the supply chain for the now infamous private label residential mortgage-backed securities (MBS’s) that contributed to the build-up of risk across complex global financial markets and led to an eventual near collapse of the global financial system in 2008. The first case will discuss the failings of mortgage origination in the U.S. mortgage system, mortgages being the financial instrument on which MBS’s are built. The second case will look at the U.S. Mortgage Electronic Registration System (MERS) that was built to facilitate the process of rapid trading of mortgages that had been securitized and turned into MBS’s. The third case explores the process of rehypothecation—when broker-dealers use client money for their own trades—a common practice in the trading of MBS’s and related financial products. The fourth case will look at what happened to the records of Lehman Brothers after the bank collapsed and the difficulties that this has created for those trying to recover their money. The final part of the paper will shift focus to the actions that need to be taken in order to build new memory systems—the next generation cyberinfrastructure—for the world’s financial institutions out of the ashes of the existing ones.

2. The Failure of Financial Memory Systems

2.1 Mortgage Origination in The U.S. Mortgage System

In the period leading up to the financial crisis, banks quickly realized that selling mortgage products to conventional, qualified borrowers alone was insufficient to meet market demand for MBS’s. To profit from the demand, banks required more mortgage transactions. As the market in MBS’s boomed, the number of the transactions, rather than the quality of the transactions, became the determinant of profit.³ To keep pace, banks invented a wide range of alternative mortgage products; at the heart of all new product offerings was the subprime mortgage.

A subprime mortgage is any loan that “exceeds the level of credit risk that government-sponsored enterprises are willing to accept for purchase.”⁴ There are many reasons why a mortgage might fall into the subprime category; chief among these is if the borrower fails to meet certain lending criteria. For example, a low credit rating (FICO score), inadequate income to service loan payments, minimal job security, negative net worth, and insufficient documentation are just a few of the characteristics common to subprime.⁵ In response to such attributes, loosened lending criteria were established in order to ‘welcome in’ a new homeowner demographic. The first lending criterion to fall afoul of this new lending policy was ‘full documentation.’

¹ Hernando de Soto, “Knowledge lies at the heart of western capitalism,” *Financial Times*, January 29, 2012.

² Ibid.

³ H. Schwartz, *Subprime Nation: American Power, Global Finance, and the Housing Bubble* (Ithaca, New York: Cornell University Press, 2009).

⁴ J. R. Barth, T. Li, T. Phumiwasana, and G. Yago, “Perspectives on the Subprime Mortgage Market,” 2008, SSRN, <http://ssrn.com/abstract=1070404> or <http://dx.doi.org/10.2139/ssrn.1070404>.

⁵ K. E. Scott and J. B. Taylor, “Why Toxic Assets are So Hard to Clean Up,” *Wall Street Journal Online*, July 21, 2009, accessed 24 March 2011, <http://online.wsj.com/article/SB124804469056163533.html>.

Full documentation generally entails the verification of both income and assets. In traditional lending practices, due diligence was paramount. Without properly drafted and signed mortgage contracts, the loan would be difficult to enforce or 'call' in the event of default. Verification of income, assets, and other personal characteristics required proper documentation in order to understand the personal and financial situation of a potential borrower, and then make effective lending decisions. In contrast to such requirements, a subprime mortgage offered little more than a signature, if that. Moreover, due diligence on these mortgages was nearly eliminated.⁶ This was an about-face from conventional lending standards that required riskier mortgages to receive greater scrutiny. One measure of this situation was the decreasing number of external reviews (i.e., due diligence) that underwriters requested. In 1995, due diligence reviewers sampled up to 30 percent of the loans in a loan pool; in 2005, arrangers were instructing due diligence firms to review only five percent.⁷ Even those responsible for checking the facts were turning a blind eye. By removing this barrier, 'limited documentation' standards allowed borrowers to qualify without demonstrating income (NISA; no income, stated assets), assets, or both (NINA; no income, no assets).⁸ At the height of reduced lending standards was the NINJA loan, where no income, no job or assets borrowers still qualified for approval. By sidestepping the need to verify anything, lenders could qualify the maximum number of borrowers in the most profitable products available and with the least hassle. Paperwork was an afterthought, particularly at the peak of the housing boom. Even the most diligent lending officer was forced to participate in the 'approve at all cost' culture. Noting the pressure on bank employees to 'sign off' on dubious loans where documentation was not present, Engel and McCoy retell the story of a Watterson-Prime employee who claimed: "...she rejected loans only to be overruled by her supervisors."⁹ Given the incentives presented to banks, the personal judgments and standards of strong-willed employees were barely recognized. On the contrary, mortgage brokers and others originating loans to feed the MBS supply chain were rewarded with large commissions and bonuses.

Decreasing standards also allowed a greater degree of fraud to go unchecked. Financial 'professionals' frequently used fake documentation and falsified statements.¹⁰ CPA Joseph T. Wells points out several tactics of fraud facilitated by forged identification and signatures, including property flipping with inflated appraisals, air loans to properties that never existed, and phantom sales where clear title houses were sold out from under the actual owner.¹¹ Each of these criminal practices required numerous fake documents, and although mortgage fraud had certainly existed long before the housing boom, the lax and permissive state of lending at the time was an enabler of financial fraud of this sort. In many cases, fraud and inefficient due diligence contributed to inadequate or faulty documentation.

Consumer awareness and education was not a priority and in some cases of predatory lending, consumers were deliberately misled;¹² a well-educated consumer would only have slowed down the origination process. This failure to communicate was particularly prevalent in interactions with subprime

⁶ K. Engel and P. McCoy, *The Subprime Virus: Reckless Credit, Regulatory Failure, and Next Steps* (New York: Oxford University Press, 2011).

⁷ Ibid.

⁸ Scott and Taylor, 2009.

⁹ Engel and McCoy, 2011.

¹⁰ Barth, 2008.

¹¹ J. T. Wells, "Mortgage Fraud: A Scourge of the 21st Century?" *CPA Journal* 79 (2009): 6-11.

¹² Federal Trade Commission, "Improving Consumer Mortgage Disclosures: An Empirical Assessment of Current and Prototype," 2007, <http://www.ftc.gov/os/2007/06/P025505MortgageDisclosureReport.pdf>.

borrowers, many of whom were unaware of the risks, costs, and long-term financial obligations that they were being encouraged to sign for.¹³ Lenders even began to steer well-qualified borrowers into subprime or other more expensive mortgage products¹⁴ often tacking on excessive fees, clauses, and conditions that were not made apparent to the average borrower.¹⁵

All of these practices resulted in poor documentation of borrowers' ability to service loans. Put simply, the documentary foundations on which the mortgages that fuelled the MBS market were very shaky indeed. This was all well and good when global demand for MBS's was strong and the game of financial musical chairs continued; however, as subsequent events revealed, it was disastrous when the music stopped.

3. The U.S. Mortgage Electronic Registration System (MERS)

A further complicating factor has been increased use of technology in the registration of mortgages. In the U.S., the transfer of mortgage titles was automated with the establishment of the Mortgage Electronic Registration Systems (MERS) in 1997.¹⁶ MERS eliminated a great deal of the paperwork involved in registering mortgages by establishing a nationwide electronic registry that, in theory, would easily keep track of who had "security interest" in a mortgage. As the volume of mortgage transactions increased in the early 2000s, the value of using MERS was apparent. Many banks, unable to produce, monitor and retain mountains of paperwork relied on MERS.¹⁷ Unfortunately, the implementation of MERS had serious flaws.

Traditionally, land and mortgage titles were held as paper records, usually housed in sturdy file cabinets at the county clerks' office.¹⁸ By law, each time a property or a mortgage title changed hands, the transaction needed to be registered with the county clerk's office. This system of registration was tenable when claims on properties and mortgage titles were between the home owner (the original borrower) and the lending institution that issued the mortgage. However, with securitization, each trade involving an MBS meant that the claim on the property and the mortgage title had to be re-registered in the name of the new owner. Keeping up with the registration of such title changes was nearly impossible for under-resourced county clerk's offices and was hugely expensive for financial institutions, as each title registration incurred a fee. MERS allowed lenders to bypass all this by registering the mortgage in its own name and established a digital repository to track the actual changes in title with each trade in the security that the original mortgage backed. Where MERS fell apart, however, was in its lack of "the thoroughness and—more importantly—the legal standing of the old system,"¹⁹ as evident in the subsequent rejection of MERS documentation by many U.S. courts.²⁰ Judges wanted to see the original signature and MERS was not designed to accommodate such requests. Tracking the ownership of a mortgage, prior to the advent of

¹³ M. Zandi, *Financial Shock: A 360 Look at the Subprime Mortgage Implosion and How to Avoid the Next Financial Crisis* (New Jersey: FT Press, 2008).

¹⁴ Barth, 2008.

¹⁵ Engel and McCoy, 2011.

¹⁶ A. Stephanie, "The Mess Gets Uglier, More Confusing," *USA Today*, n.d.

¹⁷ P. Coy, P. M. Barrett, and C. Terhuyn, "Shredding the Dream," *Bloomberg Businessweek*, 25-31 October 2010, 76-86.

¹⁸ P. Coy and J. Gittelsohn, "Mortgages Lost in the Cloud," *Bloomberg Businessweek*, 18-24 October 2010, 9-10.

¹⁹ Ibid.

²⁰ Op. Cit.

mortgage securitization, may have been feasible, but the speed and complexity of MBS trading, coupled with a paucity of accurate mortgage information, overwhelmed the system. At the height of the mortgage crisis in 2008, few people, whether bank executive or homeowner, could decipher who actually owned which mortgages.²¹

Another challenge facing MERS was whether or not, as “. . . merely a recordkeeping system,”²² it actually had the legal right to hold title and, in the case of default, foreclose on borrowers. In October of 2010, U.S. president Barack Obama came out against MERS recognition, vetoing a bill that would have confirmed the position of MERS.²³ That left many banks and mortgage holders scrambling to verify their legal claims. Ironically, banks, which previously upheld the highest standards of careful documentation, began downplaying the need for hard copy originals. Many major lenders instituted “self-imposed moratoriums”²⁴ on their foreclosure inventories at considerable cost. In other cases, mortgage holders have had to re-file for foreclosure and pay additional legal fees that the renewed process requires.²⁵ Some commentators have gone so far as to suggest that the ‘mortgage documentation crisis’ could even stall a recovery of the U.S. financial industry.²⁶

4. The Process of Rehypothecation

Rapid deleveraging in the repo markets was an important crisis contagion channel in the wake of the Lehman Brothers failure in the fall of 2008.²⁷ A “repo” is a sale of securities (i.e., collateral) combined with a simultaneous commitment to repurchase them at a later date, usually in the near term.²⁸ As Flood, Mendelowitz and Nichols explain, “a relatively simple example is a hedge fund that wants the risk and return profile of a particular security (e.g., corporate bonds) for its portfolio, but wants to boost returns by leveraging its capital. In this example, the hedge fund buys the bonds on the open market and immediately sells them into a repo transaction with its prime broker.”²⁹ The hedge fund gets the desired

²¹ Coy and Gittelsohn, 2010.

²² P. Thangavelu, “MERS Role in Foreclosure Process Under Scrutiny,” *Asset Securitization Report* 10, no. 12 (2010): 27.

²³ L. Woellert and N. Johnston, “Obama Rejects Notary Bill Amid Foreclosure ‘Caution’,” *Bloomberg Businessweek*, October 7, 2010.

²⁴ S. Lepro, “Foreclosure Restarts Face Document Doubts,” *Mortgage Servicing News* 14, no. 12 (2010): 12-13.

²⁵ Ibid.

²⁶ M. Trumbull, “Could Faulty Mortgage Paperwork Lead to a New Financial Crisis?” *Christian Science Monitor*, October 28, 2010.

²⁷ M. D. Flood, A. Mendelowitz, and B. Nichols, “Monitoring Financial Stability in a Complex World,” in *Records and Information Management for Financial Analysis and Risk Management*, ed. Victoria L. Lemieux (Frankfurt: Springer, 2012 forthcoming).

²⁸ Taub (L. Taub, “Borrowed Securities: Implications for Measuring Cross-border Portfolio Investment,” in: IFC Bulletin No. 28, IFC’s contribution to the 56th ISI Session, Lisbon, 2007 [Basle, Switzerland, Bank for International Settlements, 2010]; IMF (International Monetary Fund (IMF), “The Macroeconomic Statistical Treatment of Reverse Transactions,” Fourteenth Meeting of the IMF Committee on Balance of Payments Statistics Tokyo, Japan, October 24-26, 2001, BOPCOM-01-16, [Washington, D.C.: International Monetary Fund Statistics Department, 2001], <http://www.imf.org/external/pubs/ft/bop/2001/01-16.pdf>) and Copeland et al. (A. Copeland, A. Martin, and M. Walker, “The Tri-Party Repo Market before the 2010 Reforms,” Federal Reserve Bank of New York Staff Reports, no. 477, 2010, http://www.copeland.marginalq.com/res_doc/sr477.pdf) describe the mechanics of the repo markets in greater detail. The repo markets are very large, and there are naturally numerous variations.

²⁹ A prime broker is a specialized firm that provides a range of related services to hedge funds and other investment managers. Typical services include custody, securities settlement, tax accounting, and account-level reporting.

bonds for its portfolio, but is effectively using borrowed money to pay for them.”³⁰ The hedge fund does not receive the full value of the bonds in the “front leg” of the repo; a “haircut”, or small percentage of the value, is assessed to protect the prime broker against fluctuations in the value of the collateral. The result is one of “leveraging”, as the hedge fund can use the cash proceeds from the repo sale to purchase additional bonds. Flood, Mendelowitz and Nichols also note that it is common for the prime broker in a repo transaction to take absolute title to the collateral, which facilitates the sale of collateral by the prime broker in the event the collateral pledger fails to repurchase it as promised at the maturity of the repo.³¹

Depending on the jurisdiction and the details of the prime brokerage agreement, the collateral pledgee will have a “right to use” the collateral.³² A prime broker with a right to use may rehypothecate (re-lend) the pledger’s collateral to third parties for other purposes. For example, as Flood, Mendelowitz and Nichols observe, another hedge fund might pay to borrow the collateral to use in a short sale transaction.³³ Collateral is a scarce resource in securitization markets, so that there are strong incentives to leverage it through rehypothecation³⁴ and both the pledger and pledgee can benefit from the additional revenues generated by this reuse.³⁵

These relationships are depicted in Fig. 1, which shows both a simple repo transaction on the left and a repo involving rehypothecated collateral on the right. As Flood, Mendelowitz and Nichols note,

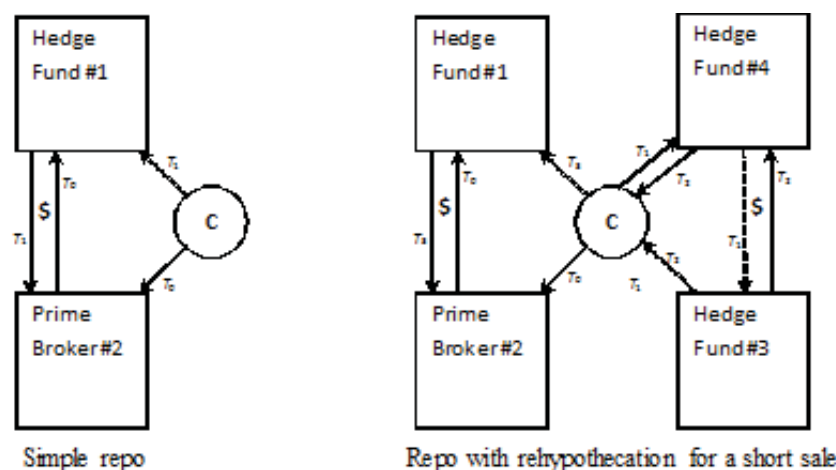


Figure 1. The impact of rehypothecation on interconnectedness³⁶

Lehman Brothers acted as prime broker for a number of large hedge funds at the time of its demise. In the example here, the hedge fund is the “collateral pledger” and the prime broker is the “collateral pledgee.”

³⁰ Flood, Mendelowitz and Nichols, 2012.

³¹ Ibid.

³² Deryugina (M. Deryugina, “Standardization of Securities Regulation: Reprehoculation and Securities Commingling in the United States and the United Kingdom,” *Review of Banking and Financial Law* 29 (2009): 253-288) describes the structure of rehypothecation transactions and related legal considerations in detail. She emphasizes the importance of the relatively lenient U.K. rules on rehypothecation in attracting prime brokerage business to London.

³³ Op. Cit.

³⁴ G. B. Gorton and A. Metrick, “Haircuts,” Working Paper 15273, National Bureau of Economic Research, Cambridge, Massachusetts, 2009, last accessed 28 August 2012, <http://nber.org/papers/w15273.pdf>.

³⁵ Deryugina, 2009, p. 257.

³⁶ Reprinted from Flood, Mendelowitz, and Nichols, 2012.

“rehypothecation occurs invisibly to the original pledger of collateral (“Hedge Fund #1” in the figure); although pledgers are aware that rehypothecation goes on, they do not generally observe when their own collateral is rehypothecated or to whom. This lack of transparency about the network of relationships played an important role in global financial crisis.”³⁷

When Lehman Brothers International Europe (LBIE) failed in London in September 2008, it had rehypothecated or commingled over \$20 billion worth of client collateral, much of which LBIE could not identify immediately.³⁸ The scramble to rapidly unwind positions in the wake of the Lehman Brothers collapse was complicated by a problem with records. Deryugina quotes from the court’s response to pledgers’ petition for information about the whereabouts of their collateral:

[I]t would be necessary to investigate particular records held by LBIE and to obtain data and records from relevant third party custodians, depositaries and other parties. ... [T]he difficulties that this process faces, not least the refusal of a number of custodians and others to comply with demands for information and that, in the meantime, the administrators are only able to call upon limited LBIE resources.³⁹

A European Senior Supervisors Group (SSG) report noted “Custody of assets and rehypothecation practices were dominant drivers of contagion, transmitting liquidity risks to other firms. The loss of rehypothecated assets and the ‘freezing’ of custody assets created alarm in the hedge fund community and led to an outflow of positions from similar accounts at other firms. Some firms’ “use of liquidity from rehypothecated assets to finance proprietary positions also exacerbated funding stresses.”⁴⁰ Similarly, a March 2009 Dear Compliance Officer letter from the UK Financial Services Authority mentions that Client Assets Sourcebook (CASS) requirements were not being followed, specifically those requiring that “a firm must keep such records and accounts as necessary to enable it—at any time and without delay—to distinguish safe custody assets/client money held for one client, from safe custody assets/client money held for any other clients and the firm’s own applicable assets/client money.”⁴¹ The letter noted how essential it is to provide an insolvency practitioner with sufficient information and cited such failures as the fact agreements and terms of business documentation was not executed with signed copies on file. “Recent firm visits,” the letter noted, “suggest that many firms do not have the appropriate trust acknowledgements in place. Where these are placed on file, we found instances where the documentation had not been executed in the name of the relevant bank or with appropriate authority on behalf of the relevant bank. Creating and operating these accounts are of paramount importance in establishing trust status for the benefit of the underlying client, the purpose of which again is only apparent on insolvency...In a period of market turbulence, we would anticipate that firms would conduct due diligence more frequently. We are reminding firms to document their due diligence.”⁴²

³⁷ Flood, Mendelowitz, and Nichols, 2012.

³⁸ Deryugina, 2009, pp. 274-275.

³⁹ Ibid., note 111.

⁴⁰ Financial Stability Board. *Risk management lessons from the global banking crisis of 2008*. Basel, Switzerland: Financial Stability Board, 2009.

⁴¹ 20 March 2009 FSA/Financial Services Authority, U.K., Dear Compliance Officer Letter, 20 March 2009, <http://www.fsa.gov>.

⁴² Ibid.

5. The Records of Lehman Brothers Post-collapse

At 7:56AM on 15 September 2008, administration orders were made in respect to each of the Lehman Administration companies. With this, Price Waterhouse Coopers (PwC) immediately assumed responsibility for the operations of Lehman Brothers International Europe (LBIE). As noted by PwC itself: “It was a monumental task.”⁴³ The administrators were faced with supporting the continued operations of some 2000 different IT systems—the financial memory systems of the firm—holding the key to what the firm owed various counterparties and how much it held in client money.⁴⁴ The task was further complicated by the fact that in the chaos and uncertainty that preceded the issuing of the administration order, many of the bank’s IT systems had already been turned off and many key IT staff had already departed. In short, the administrators faced systems of record in various states of disarray.

The complexity of LBIE’s IT architecture and the chaotic process of the firm’s demise left the administrators with many issues to sort out. Among them was the need to ensure the wind down of trading positions in the UK, which were provided by different locations and legal entities, including BarCap and Nomura.⁴⁵ Moreover, given geographic dispersion of the firm’s IT infrastructure, ownership of core IT systems was unclear and in dispute. Finally, key data that was needed to unwind positions and ascertain what money was owed to the bank’s clients was located in other geographic jurisdictions and co-mingled with the data of other entities, entities that were now competitors (i.e., BarCap and Nomura).⁴⁶

To address these issues, the administrators have had to completely restructure the firm’s IT infrastructure. Over the span of years from the collapse of LBIE to the present, the administrators have undertaken a significant rationalization of the UK-based IT infrastructure through simplification and outsourcing, and negotiated Technical Service Agreements with non-UK IT suppliers.⁴⁷ This has resulted in significant changes of the IT infrastructure from its original state and it remains unknown to what extent the original state of the infrastructure could be forensically recovered for the purposes of future research into the operations of the bank—the administrators have been very reluctant to discuss these matters given ongoing litigation. Moreover, it remains unclear what will happen to the records of LBIE once the administrators complete their work. Will the records be deposited in an archives? If so, which one? What terms and conditions will define access to the records? The answers to these questions are unknown, but for the sake of accountability to the public and future understanding of the dynamics of the financial crisis that will define this era in history, answers should be found.

6. Building New Memory Systems

The financial crisis and the Eurozone crisis that followed have drawn attention to how weaknesses in the quality and management of financial records, information, and data led to operational risks in financial institutions that prevented effective risk management. The U.S. Committee to Establish a National Institute of Finance has plainly stated that “Data management in most financial firms is a mess,” going on

⁴³ PriceWaterhouseCoopers,” Lehman Brothers International (Europe) in Administration: Joint Administrators’ Progress Report for the Period 15 September 2008 to 14 March 2009, http://www.pwc.co.uk/en_uk/uk/assets/pdf/lbie-progress-report-140409.pdf.

⁴⁴ Ibid., pp. 60-64.

⁴⁵ Op. Cit.

⁴⁶ PwC, 2009, pp. 60-64.

⁴⁷ Ibid.

to note that the absence of standard reference data, including common standardized designations for firms and their subsidiaries and for financial instruments, has hindered the way transactions are handled and recorded, and thus wasted large amounts of resources in the process of manual correction and reconciliation of millions upon millions of trades per year per firm.⁴⁸ Indeed, the need for improved data and information was specifically recognized in a Financial Stability Board (FSB) and International Monetary Fund (IMF) report on “The Financial Crisis and Information Gaps” in 2009 where it was noted that, “...the recent crisis has reaffirmed an old lesson—good data and good analysis are the lifeblood of effective surveillance and policy responses at both the national and international levels.”⁴⁹

This finding raises many questions about the characteristics of records, information, and data that produce a good result versus a bad one. Research into the conditions that contribute to good or bad results for the creation, management, and use of financial records, information, and data has implications and applications for the development of standards, best practices, and tools intended to secure a more stable financial future. With regard to governance there is a global consensus that records, information, and data management must be improved to ensure that the new institutions⁵⁰ established to govern a recalibrated post-crisis global financial system are able to provide effective financial risk analysis and management. One example of the types of initiatives being undertaken is that of the U.S. Commodity and Futures Trading Commission (CFTC), which has developed standards for describing, communicating, and storing data on complex financial products (e.g., swaps). Explaining the plan, CFTC commissioner Scott O’Malia, has said:

The data and reporting mandates of the Dodd-Frank Act place the CFTC in the centre of the complex intersection of data, finance and the law. There is a need and desire to go beyond legal entity identifiers and lay the foundation for universal data terms to describe transactions in an electronic format that are identifiable as financial instruments and recognizable as binding legal documents.⁵¹

Worldwide, members of the financial community are in the process of identifying new records, information, data requirements and standards to rebuild financial memory systems and to provide for better governance of the global financial system. Collection of data required for enhanced risk analysis and management is a huge undertaking and has given rise to many challenging discussions concerning governance of proposed records and information infrastructures, such as the discussion surrounding the need for a global identifier of all legal entities acting as counterparties to financial transactions.

Writing from a firm-centric viewpoint in an article entitled “What is Information Governance and why is it So Hard?” Gartner analyst Debra Logan argues that the root of all of our informational problems

⁴⁸ Committee to Establish the National Institute of Finance, Office of Financial Research, Landing Page, <http://www.ce-nif.org/faqs-role-of-the-nif/office-of-financial-research>, last accessed 11 August 2010.

⁴⁹ Financial Stability Board and the International Monetary Fund, “The Financial Crisis and Information Gaps,” 2009, http://www.financialstabilityboard.org/publications/r_091107e.pdf, last accessed 12 March, 2012; and Financial Stability Board and the International Monetary Fund, “The Financial Crisis and Information Gaps: Implementation Progress Report,” 2011, last accessed 12 March 2012, <http://www.imf.org/external/np/g20/pdf/063011.pdf>.

⁵⁰ There have been many new oversight initiatives and institutions established worldwide. The Dodd-Frank Wall Street Reform and Consumer Protection Act (U.S. Congress 2010) created the Financial Stability Oversight Council (FSOC) and Office of Financial Research (OFR) to monitor threats to financial stability in the U.S. The Federal Reserve Board established a new Office of Financial Stability Policy and Research. The Federal Deposit Insurance Corporation (FDIC) established a new Office of Complex Financial Institutions.

⁵¹ J. Grant, “Global derivatives lexicon edges on,” *Financial Times*, 26 June 2011, last accessed 12 March 2012, <http://www.ft.com/cms/s/0/0b170fa4-a00c-11e0-a115-00144feabdc0.html#axzz1pfGsTbOK>.

is the lack of accountability for information.⁵² Once the challenge of identifying the appropriate accountability structures for effective information governance moves beyond the traditional boundaries of the financial firm to encompass networks of international financial relationships, the ‘hard problem’ of information governance becomes acutely compounded. The issue of governance raises broad questions about the potential of policy to determine the right balance between creating and keeping records, and to ensure that the operations of financial institutions are nimble enough to respond to changing market imperatives. But good governance is in itself difficult to implement: over-regulate and operations run the risk of bogging down the delivery of financial services in overly bureaucratic regimes or encouraging regulatory arbitrage (i.e., shifting banking operations from a more regulated jurisdiction to a less regulated jurisdiction); under-regulate and the senior management, boards, shareholders, clients, and regulators of financial institutions lack the transparency required to properly assess an institution’s levels of risk.

The work of building new financial memory systems raises many technological, political, legal and ethical questions, such as to what extent governments should share their regulatory financial data with one another in the interests of global financial stability; how much of the data should be accessible to the public to promote transparency; how to keep sensitive data secure; how to harmonize metadata standards across the major global economies in the interests of creating better records and preserving them in the longer-term; and how to make well-informed decisions about what should be preserved. As difficult as these challenges are, they must be resolved if we are once again to be in a position to use the knowledge from our financial memory systems to profitably join the parts of our complex financial world together.

It is, however, not enough to focus solely on the technical aspects of recordkeeping to ensure the creation of adequate financial memory systems. Careful consideration of the human and cultural aspects of recordkeeping is also of critical importance. As the process of U.S. Mortgage origination prior to the period preceding the financial crisis illustrates, certain incentive structures will encourage poor documentation practices. Without an understanding and ability to monitor the impact of incentive structures on record making and keeping practices, the mistakes of the past are bound to be repeated with attendant detrimental effects on the world’s financial memory systems.

While new standards for records creation and metadata are key to building new financial memory systems and supporting more effective risk analysis and management across global financial markets, such memory systems must be preserved both in the usual and ordinary course of business and at times of crisis. In my view, there has been very little thinking in the global financial community about this issue. Often a forgotten aspect of the management of records, information, and data, long-term digital preservation is critical to creating the capacity for longitudinal studies of market dynamics and risk in financial institutions and financial systems. Here, it is important to make a point that many financial institutions are experiencing trouble retrieving and accessing data in as little as three to five years from the point of creation. This is due to technological obsolescence and change, as well as to a failure on the part of institutions to take measures to ensure that digital records, information, and data are created in forms that will persist over time and be properly maintained. Institutions generally have relied upon backup tapes to archive data, but this has proved to be a universally bad strategy as backup tapes are susceptible to loss and deterioration and their format makes it particularly difficult to retrieve specific and demonstrably reliable records over time. In recent years, largely in response to financial regulation (e.g.,

⁵² D. Logan, “What is Information Governance and Why is it so Hard?” *The Gartner Blog*, 11 January 2011, last accessed 12 March, 2012, http://blogs.gartner.com/debra_logan/2010/01/11/what-is-information-governance-and-why-is-it-so-hard/.

SEC rule 17-a4 1997) and high-profile litigation (e.g., *Zubulake v UBS Warburg* 2003), many financial institutions have introduced new technologies to ensure that archived records, information, and data can be reliably maintained and rapidly retrieved when needed. However, relatively little work has been done within financial institutions to address the risk factors that can lead to the long-term deterioration of reliable and authentic digital records, including efforts addressing file format obsolescence, ensuring digital records can still be read even after changes are made to data structures, as well as improving system documentation, ensuring records can still be retrieved from decommissioned systems, improving audit trails in data migration, and managing uncontrolled accumulation of records. Consequently, financial institutions and macroprudential supervisors may be unable to assume they will have access to critical information beyond a three to five-year window. Ultimately, the financial community must pay more attention to this issue. As Director of The Centre for the Investigation of Financial Electronic Records and Chair of the Open Financial Data Group Forum, I have been working to raise this issue with macroprudential supervisors and financial market participants. Much work remains to be done, however, to raise awareness and develop specific and workable solutions.

In a similar vein, macroprudential supervisors have come to understand the importance of having ready access to authentic and reliable records in the event of a financial crisis. As a result, global financial regulators are issuing new regulations requiring financial institutions to prepare “living wills” that specify how records and IT systems will be handled in the event of a collapse.⁵³ In the U.S., for example, living wills were mandated by the 2010 Dodd-Frank Act.⁵⁴ More than 100 large financial firms are required by the law to submit living wills to the Federal Reserve and the Federal Deposit Insurance Corp. by the end of 2013. Parts of the plans for nine of the biggest firms with U.S. assets have already been made public. The parts of the living will requirements that relate to records are entitled “Management Information Systems: Software Licenses: Intellectual Property” and read as follows:

xix) *Management Information Systems; Software Licenses; Intellectual Property*. Provide a detailed inventory and description of the key management information systems and applications, including systems and applications for risk management, accounting, and financial and regulatory reporting, used by the CIDI and its subsidiaries. Identify the legal owner or licensor of the systems identified above; describe the use and function of the system or application, and provide a listing of service level agreements and any software and systems licenses or associated intellectual property related thereto. Identify and discuss any disaster recovery or other backup plans. Identify common or shared facilities and systems as well as personnel necessary to operate such facilities and systems. Describe the capabilities of the CIDI’s processes and systems to collect, maintain, and report the information and other data underlying the resolution plan to management of the CIDI and, upon request to the FDIC. Describe any deficiencies, gaps or weaknesses in such capabilities and the actions the CIDI intends to take to promptly address such deficiencies, gaps, or weaknesses, and the time frame for implementing such actions.⁵⁵

⁵³ As required by the Federal Deposit Insurance Corporations (FDIC) regulation regarding resolution plans for systemically important financial institutions, (12 CFR section 381.8(c)), and, as applicable, the FDIC’s regulation regarding resolutions plans for insured depository institutions (12 CFR section 360.10(f)), each resolution plan must be divided into a public section and a confidential section.

⁵⁴ *Ibid.*

⁵⁵ 12 CFR section 360.10(f), xix, last accessed 28 August 2012, <http://www.gpo.gov/fdsys/pkg/CFR-2012-title12-vol5/pdf/CFR-2012-title12-vol5-sec360-10.pdf>.

Although the living will provisions will help to avoid the recordkeeping chaos that followed the Lehman Brothers collapse, there remain several issues. The use of the term management information systems rather than “systems of record” may not place emphasis on the type of documentation that is actually needed in a crisis. Management information systems are, by their nature, summary systems often used for risk analytics or management, not the authentic records that would be needed as proof of a claim or right of ownership of an asset in the event of a collapse. Early filings of living wills by banks with large U.S. assets suggest that these misgivings are not unwarranted. The publicly available portion of Bank of America’s living will, for example, casts the issue of preservation and continued access to its records as merely a business continuity problem.⁵⁶ In reality, it would be more appropriate if the measures were much more in line with the preservation of evidence in a forensic investigation. So far, both the regulatory requirements and the plans of financial institutions are silent as to what should happen to records during and after the process of dissolution, i.e., whether it should be necessary to forensically image systems of record, how audit trails of actions taken in respect to records systems by administrators should be captured, etc. Another issue is that each sovereign jurisdiction is issuing its own living will requirements and there is still, at present, no coordination of such provisions between global regulators. This poses a major difficulty, as most large financial institutions, such as Lehman Brothers was, operate with a global footprint. No mention has been made in the regulations about the necessity of deposit in a publicly accessible archive following administration. This leaves open the question of what will happen to the Lehman Brothers records once all the litigation has been settled and what may happen to the records of other failed financial institutions in future. Will society’s ability to understand the global financial crisis or future crises and corporate failures with major impact on society be thwarted by lack of attention or requirements to preserve the records of these firms in archives?

Weaknesses in global financial memory systems were at the heart of the global financial crisis and efforts to correct these weaknesses will be at the heart of global financial recovery. The world needs authentic and reliable memory systems so as to have the facts necessary to see risks and to protect global financial stability and growth. Research indicates that in the absence of financial memory systems, financial markets seize up and do not generate the wealth that they potentially could and that good records encourage financial trade.⁵⁷ Without strong memory systems we are likely to find our way back to the

⁵⁶ The report reads: “As required pursuant to Title I, Section 165(d), of the Dodd Frank Act and the implementing regulations (the and the FDIC’s CIDI Rule), Bank of America has developed strategies for a hypothetical resolution of its Material Entities. The Plan contemplates a resolution strategy in which Bank of America’s U.S. bank material entities (MEs), under a hypothetical failure scenario, would be resolved by placing them into FDIC receiverships. Certain assets and liabilities would be transferred to a bridge bank that would, subject to certain assumptions, emerge from resolution as a viable going concern. Bank of America’s other MEs would be wound down in an orderly manner, subject to certain assumptions. In addition, the Plan includes strategies designed to ensure continuity of certain core business lines and critical operations following the hypothetical failure of certain Bank of America entities. The strategies incorporate the importance of continued access to critical services including, but not limited to, technology, employees, facilities, intellectual property and supplier relationships.” (See, Bank of America Corporation Resolution Plan Bank of America, N.A. Resolution Plan FIA Card Services, N.A. Resolution Plan Public Executive Summary, last accessed 28 August, 2012, <http://fdic.gov/regulations/reform/resplans/boa165.pdf>).

⁵⁷ See, for example, H. De Soto, *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else* (New York: Basic Books, 2000); S. Basu, J. Dickhaut, G. Hecht, K. Towry, and G. Waymire, “Recordkeeping alters economic history by promoting reciprocity,” *PNAS* 106, no. 4 (January 27, 2009): 1009-1014; and N. Kocherlakota, “Incomplete Record-Keeping and Optimal Payment Arrangements,” *Journal of Economic Theory* 81 (1998): 272-289.

global financial crisis of 2007-2008, as recent failures of major commodities brokers suggest.⁵⁸ Where the incentives are not there within financial institutions to create and maintain such memory systems, there is a need to structure a financial system, through a balanced combination of market driven and regulatory measures, to encourage good recordkeeping. Moreover, as the UNESCO declaration on the preservation of world documentary heritage urges,⁵⁹ we need measures aimed at preserving society's financial digital heritage to ensure that it remains accessible to the public. Accordingly, access to financial memory systems, especially those in the public domain, should be free of unreasonable restrictions. The world's financial digital heritage is at risk of being lost to posterity. Just as with other records, contributing factors include the rapid obsolescence of the hardware and software that brings it to life, uncertainties about resources, responsibility and methods for maintenance and preservation, and the lack of supportive legislation. The archival community has a role to play in encouraging efforts to address these sources of risk to financial memory systems. To preserve society's financial digital heritage, measures will need to be taken throughout the financial digital information life cycle, from creation to access. Long-term preservation of financial digital heritage begins with the design of reliable and trustworthy systems and procedures that will produce authentic and stable digital objects. The many new regulatory measures aimed at strengthening global financial data standards will, to an extent, help address this issue, but much more needs to be done. For example, strategies and policies to preserve financial memory systems at times of business as usual and in times of crisis need to be developed, taking into account the level of urgency, local circumstances, available means and future projections. Cooperation among holders of copyright and intellectual property rights, and other stakeholders, in setting common standards and compatibilities, and resource sharing, will facilitate this. As with all documentary heritage, selection principles are critical, although the main criteria for deciding what digital materials to keep would be their significance and lasting cultural, scientific, evidential or other value, we will lose important knowledge about the causes of the global financial crisis, for example, if these selection criteria are made piecemeal, by individual firms or archives, or without proper consideration of how the various parts of the global financial system in general and the securitization system that brought the global financial system to its knees in 2008 operates. Moreover, these selection decisions and any subsequent reviews need to be carried out in an accountable manner, and to be based on defined principles, policies, procedures and standards explicitly designed for financial records. In addition, macroprudential supervisors need to work together to develop appropriate legal and institutional frameworks to secure the protection of the world's financial memory systems. With the proper will, this can be accomplished through the same mechanisms as such others standards as global legal entity identifiers (i.e., the G20, the ISO, and the FSB); however, awareness of the need to take action to preserve the world's financial digital heritage has to exist, alongside the political will to do so.

⁵⁸ See, for example, recent stories in the press regarding the failure of MF Global, the commodities broker, and Peregrine Financial Group (D. M. Levine, "After PFGBest, 'crisis' in commodities trading could impact everyday consumers," *The Huffington Post*, July 22, 2012, last accessed 28 August 2012, <http://www.presstv.com/usdetail/252229.html>). Both failures involved rehypothecation of client assets and inadequate trust and segregation of accounts documentation.

⁵⁹ UNESCO, "Memory of the World: General Guidelines to Safeguard Documentary Heritage," 2002, last accessed 28 August 2012, <http://unesdoc.unesco.org/images/0012/001256/125637e.pdf>.

The White House E-Mail Destruction Scandal of 2007

A Case Study for Digital Heritage¹

Myron Groover

Abstract

The 2007 controversy surrounding the e-Mail transmission and retention practices of the George W. Bush Jr. White House marked the beginning of a pivotal episode which saw the idea of digital preservation itself pushed to the forefront public consciousness. It culminated in a lengthy and conspicuously inconclusive lawsuit which saw both the National Archives and Records Administration and the Executive Office of the President as co-defendants; a lawsuit which was eventually resolved out of court when the White House changed political hands. The events concerned, and the extrajudicial resolution which effectively closed the case, have much to tell us about the politics—and laws—which impact the present and future state of digital heritage in the United States and worldwide. I present here a brief overview of the scandal, the contexts surrounding it, and some lessons we might hope to learn from the events concerned.

Author

Myron Groover is an archivist at the Vancouver Holocaust Education Centre and a member of BCCLA's information Policy Committee. He holds an MA (hons) in History from the University of Aberdeen and recently completed his MAS and MLIS at the University of British Columbia. During the course of his academic and professional work with public libraries, archives, and other heritage institutions, he has developed a keen interest in the ways archival and library theory (and practice) influence and relate to the shaping, maintenance, and legitimization of discourse within contemporary and historical nation-states.

1. Introduction

The records of governmental bodies, particularly at higher judicial, legislative and executive levels, are as invaluable as they are irreplaceable. They are the living records of a society's life and are the rightful heritage not only of that society's citizens but indeed of the entire world—a world which will one day look to them in its attempts to make sense of its past just as we look to the records of past governments today. Imagine, for example, what our understanding of Nixon's presidency might look like without the Watergate tapes; now imagine how immeasurably the digital revolution has changed the scale of what's at stake since then. The digital arena enables us to preserve more and better records than has ever been possible in human history—but it also presents us with similarly unprecedented opportunities to tamper with, destroy, or even lose those records outright.

The 2007 controversy surrounding the e-Mail transmission and retention practices of the George W. Bush Jr. White House was an important watershed moment for digital heritage. While it was by no means the first time a United States presidential administration found itself in hot water over electronic record-keeping, it did mark the beginning of a particularly pivotal episode which saw these weighty issues—and the very idea of digital preservation itself—pushed to the forefront of public consciousness. It culminated

¹ This work represents a distillation and maturation of my work to date regarding these events. I wish to thank Jason R. Baron in providing comments on an earlier draft of this paper; the views expressed herein are entirely my own, however, as are any errors or omissions.

in a lengthy and conspicuously inconclusive lawsuit which saw both the National Archives and Records Administration and the Executive Office of the President as co-defendants; a lawsuit which was eventually resolved out of court when the White House changed political hands.² The events concerned, and the extrajudicial resolution which effectively closed the case, have much to tell us about the politics—and laws—which impact the present and future state of digital heritage in the United States and worldwide.

This paper will give (1) a brief synopsis of the political and administrative contexts surrounding the controversy, (2) a limited chronology of the events which precipitated it and brought it to resolution, and (3) an investigation of the ongoing legal and ethical issues highlighted by the incidents concerned. All these components will focus on the inadequacy of current safeguards in the United States to protect and preserve digital heritage artefacts generated by powerful administrative bodies—ultimately, as a candid investigation of the existing legal and administrative frameworks will show, it is too easy for attitudes and practices regarding digital heritage preservation at higher levels of government to be dictated by political will rather than a spirit of compliance and accountability. This situation is untenable, so my presentation will conclude with (4) a series of questions and recommendations which may serve to inform the development of future decision-making regarding the digital heritage of higher governmental bodies both in the United States and around the world.

2. Background and Context

Controversy surrounding the e-mail practices of the federal government did not begin in 2006 when the first public murmurings that all might not be well with the Bush administration's electronic record-keeping began to surface.³ Indeed, as prior scholarship has pointed out, legal wrangling over the statutory obligations which surround the preservation and destruction of federal e-mail dates back to at least the late 1980s when the grounds were laid for what would be the first of many rounds of litigation surrounding the PROFS e-mail system in use by the National Security Council and other federal bodies at the time. The scandals of 2007 took place not only in light of a long and complex litigative legacy but also in the context of the unprecedented rise of the internet itself as a phenomenon—the electronic landscape of the second George W. Bush administration was markedly different than that which predominated in Oliver North's day, but many of the actual blunders involved in the Bush scandal bore a remarkable similarity to the events which gave rise to the original PROFS case.⁴

From a statutory viewpoint, there are two key pieces of federal legislation in the United States which treat very directly on the subject of presidential records. The most notable is the Presidential Records

² It is significant, if perhaps not unusual, that the circumstances of the scandal first became known only coincidentally. The initial evidence came to light as a side-note of unrelated investigations into the Bush administration – one regarding the Plame Affair and one regarding the dismissal of US Federal attorneys for political purposes – and without these broader scandals the events discussed here would have taken an entirely different shape.

³ For more information on the circumstances surrounding the early days of the scandal and the media coverage thereof, see Media Matters Research Team, "Media Largely Ignored Fitzgerald Revelation that White House May Have Destroyed Emails," Media Matters for America, accessed August 31, 2012, <http://mediamatters.org/research/200602020012>.

⁴ Jason Baron, "The PROFS Decade: NARA, E-mail, and the Courts," in *Thirty Years of Electronic Records*, ed. Bruce L. Ambacher (Lanham: Scarecrow Press, 2003), 105-37.

Act,⁵ first passed in 1978 and amended three times subsequently, which holds as a central tenet that “[t]he United States shall reserve and retain complete ownership, possession, and control of Presidential records”⁶ and that:

...[t]hrough the implementation of records management controls and other necessary actions, the President shall take all such steps as may be necessary to assure that the activities, deliberations, decisions, and policies that reflect the performance of his constitutional, statutory, or other official or ceremonial duties are adequately documented and that such records are maintained as Presidential records pursuant to the requirements of this section and other provisions of law⁷

In 2002, President Bush amended the law by executive order to severely curtail the public *availability* of records preserved because of the legislation;^{8,9,10} he did nothing, however, to alter the fundamental requirement that presidential records be maintained in the first place. Any destruction of e-Mails which fall under the purview of “Presidential Records” as outlined above—whether deliberately, at the whim of staffers or party leadership; or unintentionally through neglect and incompetence—constitutes a clear violation of the PRA.

Another significant law influencing the ongoing discourse around defining the norms of presidential record-keeping is the Hatch Act of 1939, which was originally passed to curb alleged corruption and political patronage tied up with various New Deal programs. The law has since become particularly important for its prohibition on the use of governmental resources for partisan political activities. By blurring the lines between party-political maneuvering and official activity relating to the business of government, officials can either seek to pass off party business as federal business worthy of federal resources—or, on the opposite end of the spectrum, argue that communications which mention party business are out with the sphere of federal record-keeping legislation.

⁵ United States, Congress, 1978, Presidential Records Act, *United States Code, Title 44*, accessed 7 April 2010; available from http://en.wikisource.org/wiki/Presidential_Records_Act_of_1978.

⁶ *Ibid.*, §2202

“Presidential Records” are here defined as:

“... documentary materials, or any reasonably segregable portion thereof, created or received by the President, his immediate staff, or a unit or individual of the Executive Office of the President whose function is to advise and assist the President, in the course of conducting activities which relate to or have an effect upon the carrying out of the constitutional, statutory, or other official or ceremonial duties of the President ... includ[ing] any documentary materials relating to the political activities of the President or members of his staff, but only if such activities relate to or have a direct effect upon the carrying out of... official or ceremonial duties of the President.” (§2201)

“Documentary materials” are defined as:

“... books, correspondence, memorandums, documents, papers, pamphlets, works of art, models, pictures, photographs, plats, maps, films, and motion pictures, including, but not limited to, audio, audiovisual, or other electronic or mechanical recordings” (§2201)

⁷ *Ibid.*, §2203

⁸ George Bush, “Source Material: Executive Order 13233 [Federal Register Vol. 66, No. 214, November 5, 2001] Further Implementation of the Presidential Records Act,” *Presidential Studies Quarterly* 32, no. 1 (March 2002): 185-89, accessed August 31, 2012, <http://www.jstor.org/stable/27552373>.

⁹ See also Bruce Montgomery, “Source Material: Nixon’s Ghost Haunts the Presidential Records Act: The Reagan and George W. Bush Administrations,” *Presidential Studies Quarterly* 32, no. 4 (December 2002): 789-809, accessed August 31, 2012, <http://www.jstor.org/stable/27552442>.

¹⁰ See also Bruce Montgomery, “Presidential Materials: Politics and the Presidential Records Act,” *The American Archivist* 66, no. 1 (Spring/Summer 2003): 102-38, accessed August 31, 2012, <http://www.jstor.org/stable/40294220>.

3. The Controversy – A Narrative Overview

The scandal itself fundamentally concerned inconsistencies in the Bush Administration's recordkeeping practices which resulted in haphazard archival retention of both official and unofficial White House e-mail. It also concerned the subsequent attempts of White House staffers to illegally destroy e-Mails for a variety of reasons. These practices and processes took place far from public scrutiny and governmental oversight, and were subsequently found to be in violation of several federal statutes. The matter gave rise to public outrage at both the politicization of the Executive Branch of the United States federal government and at the seeming audacity of the Executive in seemingly behaving as if it was pre-eminent among the branches of government and therefore not subject to the oversight or separation of powers set forth in United States law.

The scandal has its origins in the early years of the 21st century, but it was in 2004 that the scandal as a phenomenon began to take shape. It was the third year of an ongoing grand jury investigation into the Plame Affair whereby Richard Armitage and others made public the identity of CIA covert agent Valerie Plame as a political reprisal for her husband's alleged opposition to Bush administration interests. It was during the indictment and trial of President Cheney's then-Chief of Staff Lewis "Scooter" Libby for his involvement in the incident that the first public mention of inconsistencies in the Bush White House's e-mail practices came to light.¹¹ In a letter to Libby's attorneys from special counsel Patrick Fitzgerald, who was responsible in large part for prosecuting the case, we see the somewhat understated remark that:

In an abundance of caution, we advise you that we have learned that not all e-mail of the Office of Vice President and the Executive Office of the President for certain time periods in 2003 was preserved through the normal archiving process on the White House computer system.¹²

As time went on, these discrepancies began to be subjected to more intense scrutiny. In 2007, during an investigation of the White House for its involvement in illegal Department of Justice efforts to politicize the hiring and firing of United States Attorneys, the Bush administration was compelled to reveal in response to official subpoenas that it could not provide many of the e-Mails requested by the prosecution because these E-Mails had not been sent or received using the official White House E-Mail servers. This admission eventually resulted in the discovery that a significant percentage of the E-Mail sent by White House staffers was being hosted by internet domains registered to such overtly partisan entities as the Republican National Committee or "Bush-Cheney '04 Inc."¹³ The most widely-used of these domains became a focal point for a subsequent investigation of the Bush White House by the House Judiciary

¹¹ For a parallel case, see Nick Juliano, "Ethics Watchdog Accuses Education Department of Illegal E-Mail Use," *The Raw Story*, accessed August 31, 2012, http://rawstory.com/news/2007/Ethics_watchdog_accuses_Education_Department_of_0516.html.

¹² Patrick Fitzgerald to Messrs. Jeffress, Wells and Tate, "RE: United States vs. L. Lewis Libby," Letter, January 23 2006, in *CREW Lawsuit Exhibits*, Citizens for Ethics and Responsibility in Washington, accessed August 31, 2012, http://www.citizensforethics.org/files/Exhibits_1.pdf.

¹³ For registration information regarding these domains, see <http://whois.domaintools.com/georgewbush.com>, all accessed August 31, 2012, <http://whois.domaintools.com/gwb43.com> and <http://whois.domaintools.com/rnchq.org>.

Committee.¹⁴ The matter was further compounded by the fact that, as it subsequently became clear, the White House's record-keeping procedures in place for even official e-mail sent on federally-owned servers were well below standard.

During the course of several investigations, a number of allegedly-destroyed E-Mails explicitly discussing both policy *and* politics came to light. The mixed nature of these e-mails' content brought the Hatch Act and Presidential Records Act, and the tension between the two, into the spotlight. By arguing that the unrecoverable e-mails deleted or not preserved on federal servers—or sent via entirely external domains in the first place—were exclusively used for partisan purposes and politicking rather than the business of government, the White House could offer some justification for certain e-mail having gone missing. This justification wasn't entirely enough to overshadow the statutory requirements of the Presidential Records Act, however, and in acknowledgement of that fact the White House's Deputy Press Secretary Scott Stanzel was compelled to announce that non-federal domains had indeed been used for official government business by staff, that “the White House has not done a good enough job overseeing staff using political e-mail accounts to assure compliance with the Presidential Records Act,”¹⁵ that “...some official E-Mails have potentially been lost,” and that the White House would take steps to recover the lost E-Mails. Stanzel stopped short of admitting any deliberate wrongdoing, maintaining that the confusion stemmed from an ‘unclear’ policy on E-Mail within the White House.¹⁶

Interpretation of the above events was undoubtedly influenced by the publication in April 2007 of a highly critical report entitled *Without a Trace: the Story Behind the Missing White House E-Mails and the Violations of the Presidential Records Act*.¹⁷ The report was authored by Washington-based civilian watchdog CREW (Citizens for Responsibility and Ethics in Washington), a whistleblower organization whose stated purpose is to “promote ethics and accountability in government and public life by targeting government officials—regardless of party affiliation—who sacrifice the common good to special interests.”¹⁸ It laid out in no uncertain terms what CREW perceived to be egregious law-breaking and deliberate misinformation on the part of the White House. The report caught the attention of the press and Congress with its projected figure of five million potentially-lost E-mails—a number which White House Press Secretary Dana Perino referenced directly the very next day with the words “I wouldn't rule out that there were a potential 5 million emails lost.”¹⁹ This seeming admission of culpability was tempered somewhat when months later a White House spokesperson named Tony Fratto alleged that “to the best of

¹⁴ Naomi Steiner, “CREW Asks for House Investigation into White House Violations of Presidential Records Act,” Citizens for Responsibility and Ethics in Washington, accessed August 31, 2012, <http://www.citizensforethics.org/node/25822>.

¹⁵ Tom Hamburger, “Key Bush Aides' E-Mail May Be Lost,” *Los Angeles Times*, April 12, 2007. accessed August 31, 2012, <http://articles.latimes.com/2007/apr/12/nation/na-emails12>.

¹⁶ Sheryl Stolberg, “Bush Advisers' Approach on E-Mail Draws Fire,” *New York Times*, April 12, 2007. accessed August 31, 2012, <http://query.nytimes.com/gst/fullpage.html?res=950CE1DA123FF931A25757C0A9619C8B63>.

¹⁷ CREW, “Without a Trace: the Story Behind the Missing White House E-Mails and the Violations of the Presidential Records Act,” April 12, 2007, Citizens for Ethics and Responsibility in Washington, accessed August 31, 2012, <http://www.scribd.com/doc/132187/Without-a-Trace-The-Missing-White-House-Emails-and-the-Violations-of-the-Presidential-Records-Act>.

¹⁸ CREW, “About CREW,” Citizens for Responsibility and Ethics in Washington, accessed August 31, 2012, <http://www.citizensforethics.org/about>.

¹⁹ Anonymous, “CNN: White House Won't Deny Lost E-Mails,” YouTube, accessed August 31, 2012, <http://www.youtube.com/watch?v=ryCZ9eZrvT8>.

what all the analysis [sic] we've been able to do, we have absolutely no reason to believe that any emails are missing; there's no evidence of that."²⁰

The situation was escalated by lawsuits filed in the autumn of 2007 by two watchdog organizations—first the National Security Archive²¹ and subsequently by the aforementioned Citizens for Responsibility and Ethics in Washington.²² Both filed suit not only against the Executive Office of the President but also against Allen Weinstein and NARA, alleging that the latter parties were complicit in the scandal since they had failed to take effective action to monitor the goings-on of record-keeping practices in the Bush White House or to raise any alarm about what was (depending on your interpretation) either a serious ethical breach or an alarming incidence of technological incompetence.

As the lawsuits were gathering steam, the entire affair came to a head in early 2008 at a dedicated hearing before the House Committee on Oversight and Government Reform which saw Bush Administration IT officials give their accounting of what had taken place and why precisely so many e-mails were missing. The public record of this hearing and the interrogatories which surrounded it provide one of the most valuable resources for studying the events concerned. The Hatch Act and Presidential Records Act were discussed extensively, but perhaps more significant were the statements of IT officials regarding their genuine technical difficulties in preserving the White House's e-Mails. Far from presenting a picture of malfeasance and conspiracy, the White House's CIO Theresa Payton testified to a challenging work climate of vast workloads, insufficient budgetary staff resources, and a rapidly-shifting technological landscape which confounded the good-faith efforts of White House IT staff to preserve e-mail in accordance with the provisions of the PRA.²³ The particulars of this testimony were broadly supported by a written interrogatory from Steven McDevitt, a former administration IT official, although McDevitt went into much greater technical detail on the specifics of White House information policy. McDevitt's interrogatory did contrast with Payton's testimony in one important particular, however—rather than putting the failure to adequately preserve White House e-mail down to overwork and finite resources, he instead painted an extremely unflattering picture of incompetence and inadequacy among OCIO staff.²⁴

This hearing added fuel to the ongoing litigation, which wound its way through the courts inconclusively during most of 2008 and the early part of 2009. After a lengthy period of discovery, the case was eventually put on hold some two months after the swearing-in of the Obama administration, which had used transparency and openness as campaign watchwords.²⁵ During this time, the new

²⁰ CREW, "A New Claim from the Bush Administration: 'We Have Absolutely No Reason to Believe That Any Emails Are Missing'," Citizens for Responsibility and Ethics in Washington, January 17, 2008, accessed August 31, 2012, <http://www.citizensforethics.org/node/30806>. This statement was questioned the next day in a letter from Henry Waxman, chairman of the House Judiciary Committee, pointing out that a 2005 report had identified nearly 500 days where the White House registered no E-Mail traffic at all.

²¹ Meredith Fuchs and Tom Blanton, "Archive Sues to Recover 5 Million Missing White House E-mails," National Security Archive, accessed August 31, 2012, <http://www.gwu.edu/~nsarchiv/news/20070905/index.htm>.

²² Anne Weismann and Melanie Sloan, "Citizens for Responsibility and Ethics in Washington vs. Executive Office of the President," Lawsuit, September 25 2007, National Security Archive, accessed August 31, 2012, http://www.gwu.edu/~nsarchiv/news/20080417/092507_CREW_Complaint.pdf.

²³ U.S. House of Representatives Committee on Oversight and Government Reform, "The Electronic Records Preservation at the White House Hearing before the Committee (Serial No. 110-80)", U.S. Government Printing Office, February 26, 2008, accessed August 31, 2012, <http://www.fas.org/sgp/congress/2008/electronic.pdf>.

²⁴ Ibid.

²⁵ Judicial Watch, "Obama Promises Transparency, Openness," Judicial Watch, November 10, 2007, accessed August 31, 2012, <http://www.judicialwatch.org/blog/2008/nov/obama-promises-transparency-openness>.

administration began pursuing the restoration of the supposedly-lost E-Mails fairly aggressively, and after several months in abeyance the suit was finally settled out of court in December 2009.²⁶ The defendants agreed to restore some 94 days' worth of previously-missing E-Mails and to transfer them to NARA, and to further ensure that all back-up tapes relating to the period in question were retained for a full twelve years. The final tally indicated that some number approaching 23 million E-Mails had been destroyed or lost by the Bush White House,²⁷ a far cry from the figure of five million which was bandied about in the early days of the scandal.

4. Interpretation and Aftermath

As the investigations progressed, two broad narratives emerged regarding the circumstances of the scandal. The first, endorsed by administration officials, is perhaps best encapsulated by the testimony of Theresa Payton and by the following response given by Scott Stanzel to a particularly pointed question regarding the Presidential Records Act on the day the initial CREW report was published:

Well, technology has certainly advanced. We live in a new time. This is just the second administration who's actually had E-Mail. This is the first administration who has dealt with the ubiquity of 24/7 communications in the form of BlackBerrys. So it is always on. The White House policy actually has been improved. We've strengthened that in policy, clarified it for staff so they understand how to avoid violations of the Hatch Act, while at the same time adhering to the Presidential Records Act ... There are official business emails, there are political business emails, and then there is also this gray area. And that's where employees have to make a judgment. And some employees, out of an abundance of caution, could have been sending official business emails on their RNC political account.²⁸

This narrative centers around the technological and policy dimensions of electronic preservation, holding that any gaps in the official record are due to both policy confusion among staffers and to technical inadequacies on the part of the White House's information infrastructure which compromised e-mail retention in some way. Opinion on the part of the IT personnel involved is mixed, as we have seen; there is by no means a consensus on whether this came to be through incompetence or misfortune, but at its core this narrative does not emphasize any malfeasance or deliberate wrongdoing.

The second dominant narrative, as espoused by CREW through publications such as *Without a Trace* and through the CREW-NSA lawsuit, holds that the loss of so many e-mails cannot be ascribed to either incompetence or technological inadequacy—instead, proponents say, the scandal is evidence of a pernicious and long-enduring culture of obfuscation and callousness regarding federal recordkeeping statutes on the part of the Bush White House.

²⁶ Dan Eggen, "Groups Announce Settlement in Missing Bush E-Mails Case," *Washington Post*, December 14, 2009. accessed August 31, 2012, <http://voices.washingtonpost.com/44/2009/12/groups-announced-settlement-in.html>.

²⁷ Cable News Network, "Millions of Bush Administration E-Mails Recovered," *CNN*, December 14, 2009. accessed August 31, 2012, <http://www.cnn.com/2009/POLITICS/12/14/white.house.emails/index.html>.

²⁸ Paul Kiel, "White House: 'We Live in a New Time'," Talking Points Memo, accessed August 31, 2012, <http://tpmmuckraker.talkingpointsmemo.com/archives/002993.php>.

While neither narrative eventually prevailed in the courts, frustration on the part of NARA increased as the Bush Administration drew to a close. Allen Weinstein, the Archivist of the United States, sent White House Counsel Fred Fielding a letter pleading that the latter commence E-Mail restoration efforts.²⁹ A meeting was subsequently arranged between National Archives and Records Administration representatives and White House staff—largely to give the latter a chance to explain to the former why they believed so many E-Mails were missing and what could be done to recover them if that were possible. According to leaked minutes of the meeting, NARA “... wanted a complete set of Bush emails in a format that we could accept into ERA”—but the organization had no power to force the White House to comply or facilitate this.³⁰ This frustration was voiced in an internal NARA memo from Counsel Gary M. Stern to Allen Weinstein, stating that “...we still have made almost zero progress in actually moving ahead with the important and necessary work ... even our rather simple and mundane request[s] ... [have] lain dormant for months.”³¹

Whether because of malfeasance or technological factors (or a combination of both) NARA encountered real difficulty in its efforts to safeguard and preserve the electronic records of the Bush White House, immersed as it was in the very involved process of designing and implementing its flagship ERA (Electronic Records Archives) program.³² These problems were compounded by NARA’s statutory mandate, which leaves it with little power to force compliance or oversight on powerful federal agencies. The end result was a serious breach in electronic record-keeping at the very highest levels of the United States Government whose impacts are still being felt today.

The White House itself was left somewhat shamefaced by the controversy, but in spite of a fairly broad consensus that the White House had not sufficiently discharged its responsibilities under the Presidential Records Act no criminal case was ever brought against anyone as a result of the scandal. Congress did not take any action to compel the White House to rectify its mistakes by restoring lost E-Mails, and the whole incident was generally treated by Congress as an aside to the much more intense controversy over the politicization of U.S. Attorney firings by the Department of Justice. The lawsuit against the EOP was eventually settled—albeit after a tortuous litigative procedure that would have bankrupted any but the most well-funded NGOs—but the issues raised by the controversy are far from resolution. Not least among these are the lingering questions as to the robustness of federal record-keeping practices and the information architecture which supports them—and questions about the ability of these practices and systems to support a culture of heritage preservation and accountability.

²⁹ Allen Weinstein to Fred Fielding, Letter, May 12 2007, in *National Security Archive Primary Documents*, National Security Archive, accessed August 31, 2012, http://www.citizensforethics.org/files/Exhibits_1.pdf.

³⁰ NARA, “Notes of May 21, 2007 Meeting with White House on Email Issues,” National Security Archive, accessed August 31, 2012, [http://cdn.preterhuman.net/texts/government_information/Notes of May 21, 2007 Meeting with White House on Email Issues.pdf](http://cdn.preterhuman.net/texts/government_information/Notes%20of%20May%2021,%202007%20Meeting%20with%20White%20House%20on%20Email%20Issues.pdf)

³¹ Gary Stern to Allen Weinstein, “Bush 43 Transition, et. al.,” Memorandum, September 5 2007, in *National Security Archives Primary Documents*, National Security Archive, [http://www.gwu.edu/~nsarchiv/news/20080417/09052007 Stern-Weinstein memo.pdf](http://www.gwu.edu/~nsarchiv/news/20080417/09052007%20Stern-Weinstein%20memo.pdf) (accessed August 31, 2012)

³² See the United States Budgets, where for the years from 2002-2007 the only line item consistently mentioned in NARA’s written appropriations summaries was ERA. Searchable versions of past federal budgets can be found at: Executive Office of the President, “Budget of the United States Government: Browse,” Government Printing Office, <http://www.gpoaccess.gov/usbudget/browse.html> (accessed August 31, 2012).

5. Analysis and Recommendations

For an analysis of the record-keeping proficiency of the Bush White House, we might look to the ARMA International Draft Maturity Model as a frame of reference. We might reasonably hope that the Executive Branch³³ of the most powerful government on the planet would meet or approach ARMA's highest level of standards—5—but the facts which came to light as a result of the scandal and ensuing litigation have revealed that the core competency of the White House was closer to ARMA's level 2 in most areas, or perhaps even 1 (the lowest possible level) in terms of Transparency,³⁴ Availability,³⁵ Integrity³⁶ and Retention.³⁷ This analysis stands whether or not the errors in electronic record-keeping were deliberate or accidental, but we should not lose sight of the GARP Principle of Transparency:

The processes and activities of an organization's record-keeping program shall be documented in an understandable manner and be available to all personnel and appropriate interested parties ... It is in the best interest of every organization, and of society in general, that all parties clearly understand [that] [1] The organization conducts its activities in a lawful and appropriate manner. [2] The record-keeping system accurately and completely records the activities of the organization. [3] The record-keeping system is itself structured in a lawful and appropriate manner. [4] Activities conducted to implement the record-keeping program are conducted in a lawful and appropriate manner.³⁸

Whatever political position we take on the exigencies of the scandal, the White House does not come out looking especially favourable. Ultimately, the issue is not whether or not the failures of the Bush Administration in preserving its e-mail were down to malfeasance or circumstance—it is whether those failures are *acceptable* to us. More broadly, thinking along such lines invites us to examine our existing statutes and attitudes to evaluate their ability to support the culture of record-keeping we would like to see.

What effective deterrents, professional or legal, are in place at the very highest level of society and governance to ensure that these principles are being applied? In this instance, the immediate real legal impact of the incident was reduced to a two-year litigious standoff (at the expense of the American

³³ Due largely to a series of political circumstances and unprecedented interpretations and re-workings of the mechanics of government, the Executive Branch was certainly the most immediately influential arm of American government during virtually the entirety of Bush's two terms in office, which may mitigate or exacerbate the implications of his administration's cavalier behavior regarding Records Management.

³⁴ ARMA, "ARMA International Draft Maturity Model," Association of Records Managers and Administrators International, accessed August 31, 2012, http://www.armacanada.org/documents/GARP_MaturityModelchart.pdf. "It is virtually impossible to obtain information about the organization or its records in a timely fashion ... There is no emphasis on transparency. It is not clear how the organization operates."

³⁵ Ibid. "Records are not readily available when needed ... Record inventories are practically non-existent ... Discovery is difficult since it is not clear where information resides or where the copy of the record is"

³⁶ Ibid. "There are no systematic audits, and no defined process for showing the origin and authenticity of a record. Various organizational functions use ad hoc methods to demonstrate authenticity and chain of custody ... but their trustworthiness cannot be easily guaranteed."

³⁷ Ibid. "There is no current documented records retention schedule. Rules and regulations that should define retention are not identified or centralized ... In the absence of retention schedules, employees either keep everything or dispose of records based upon individual rather than organizational needs."

³⁸ ARMA, "Generally Accepted Record-keeping Principles (GARP) : Transparency," Association of Records Managers and Administrators International, accessed August 31, 2012, <http://www.arma.org/garp/>.

taxpayer and the financial supporters of the NGOs who brought the suit) which could only reach some modicum of resolution when the Executive Branch itself changed hands and ceased actively obstructing the course of the investigation, resulting in an out-of-court settlement.

For answers, perhaps we can look to the Universal Declaration of Archives ratified by the ICA in 2011. It provides a fairly unambiguous and encouraging blueprint for confronting just such issues. If we are to avoid similar incidents in the future, we must put the Declaration into practice—thankfully, it tells us exactly what we must work towards, a climate where:

...appropriate national archival policies and laws are adopted and enforced; the management of archives is valued and carried out competently by all bodies, private or public, which create and use archives in the course of conducting their business; adequate resources are allocated to support the proper management of archives, including the employment of trained professionals; archives are managed and preserved in ways that ensure their authenticity, reliability, integrity and usability.³⁹

In light of this, it is abundantly clear that relying on litigation to ensure compliance is inadequate. Such processes can be profoundly undemocratic, relying heavily on financial resources to effectively buy outcomes, but are also extremely inefficient when one party to a suit is powerful enough to be consistently uncooperative and to delay the resolution of the suit indefinitely. Nor can we legitimately expect large federal bodies to self-regulate.

Instead, we might ask hard questions of ourselves about our willingness—as professionals, and as a society—to pursue alternate routes toward ensuring compliance. Most crucial of all is the adequate funding of bodies like NARA, which are our first line of defense in the battle to protect our digital heritage. Such bodies must also be innovative—as NARA indeed is—in exhaustively documenting financial or material impediments to their efforts in this vein, so that it can be demonstrated that archivists and other heritage keepers have truly done all they can to facilitate good practices in electronic record-keeping. It is a matter of efficiency as much as anything else; in the case of the White House e-mail scandal, many e-mails were eventually recovered (electronic media seldom being so ephemeral as users think), but how much easier would it have been if adequate retention practices had been in place from the beginning?

The elephant in the room, perhaps, is prosecution. Ultimately, if large-scale legislative measures like the Presidential Records Act are to prevent our digital heritage from being compromised, we cannot shy so readily away from criminal proceedings against those found to be responsible for lapses like those which gave rise to the Bush e-mail scandal. This is not whatsoever to say that there should be no room for negotiation or diminished responsibility in light of circumstance, but our unwillingness at a societal level to see the mighty in court over such matters cannot continue if we are to remain credible.

If it is expected these archival principles will be adhered to by business leaders and governments, how can we vouch for the universality of our methodology if there is a tacit admission that some are beyond the reach of approbation for violating its tenets? Do our generally-accepted ways of defining records, as in the case of the legitimate conflict between the Presidential Records Act and the Hatch Act, truly lend themselves to regulation and a need to protect against conflicts of interest?

³⁹ ICA, “Universal Declaration on Archives,” International Council on Archives, accessed August 31, 2012, <http://www.ica.org/download.php?id=2027>.

In the case of electronic heritage preservation these questions are especially poignant. Ultimately it may be that the world *needs* to see a high-profile case around these issues—visible neglect on the part of powerful governmental bodies in capturing, maintaining, and ultimately retaining their electronic legacy, whether due to malfeasance or technical inadequacy—successfully brought to conviction in a court of law. We still have a chance to shape the debate around electronic heritage preservation, if we are willing.

None of these measures can stand alone, however. As the Universal Declaration on Archives so clearly indicates, we need to continue working towards an *integrated* culture of reverence and respect towards our emerging digital heritage—and that goes far beyond shaming or punishing people who break the law. Instead, we should strive for a society where principles like those which inspire the Declaration—and this conference—are so generally accepted that they require neither punishment nor praise to maintain.

An archivist can dream, after all.

The Perfect Archival Storm

The Transfer of Electronic Records from the G.W. Bush White House to the National Archives of the United States

Kenneth Thibodeau

Abstract

The National Archives and Records Administration (NARA) of the U.S. has experienced several important growth spurts in its holdings of electronic records during the last two decades. Sudden, exponential increases in transfers of electronic records have principally been an artefact of the Presidential Records Act, which provides that as soon as a president leaves office all extant presidential records of that administration become the legal responsibility of the Archivist of the United States. The article describes these growth spurts and recounts how NARA has dealt with them.

Author

Dr. Kenneth Thibodeau is an internationally recognized expert in electronic records and digital preservation. A senior guest scientist in the Information Technology Laboratory of the National Institute of Standards and Technology of the U.S., he previously directed the Center for Advanced Systems and Technology and the Electronic Records Archives Program at the National Archives and Records Administration in Washington. He also served as the Chief of Records Management at the National Institutes of Health and directed the Department of Defense Records Management Task Force, leading the development of the world's first standard for records management software. Fellow of the Society of American Archivists, Thibodeau won the Emmett Leahy Award and a Lifetime Achievement Award from the Archivist of the United States.

1. Historical Background

Electronic records initially attracted the attention of the National Archives and Records Administration (NARA) of the United States starting in the 1960s. After several years of analysis and organization, the National Archives first accessioned electronic records appraised as having permanent value in 1970. NARA continued to develop capabilities for processing and preserving electronic records, focusing its efforts in a single organizational unit, the Machine-Readable Archives Division until this growth was aborted in the early 1980s under a general reduction of federal government programs during the administration of President Reagan. Reduced from division to branch level, losing significant numbers of staff, and stripped of internal technical capabilities, the Machine-Readable Archives unit stagnated through most of the eighties. By the end of the decade, however, NARA management decided it needed to increase attention to electronic records. At the end of 1988, it raised the status of the machine-readable archives unit back to division level, renaming it the Center for Electronic Records. Initially, the Center had no more resource than the predecessor branch; however, NARA did transfer responsibilities for all archival processes related to electronic records to the Center and gradually increased its staffing and budget over several years.¹

¹ Thomas E. Brown, "History of NARA's Custodial Program for Electronic Records: From the Data Archives Staff to the Center for Electronic Records, 1968-1998," in *Thirty Years of Electronic Records*, ed. Bruce Ambacher (Lanham, Maryland: The Scarecrow Press, 2001): 1-24.

With the unification of responsibilities for electronic records, the concentration of expertise, and increasing resources, the Center was able to make considerable progress. It developed rigorous and in-depth procedures for appraisal, grounded in well-established archival law, theory, and practice, but refined for the special characteristics of electronic records.² In less than five years, it doubled the number of series of electronic records appraised as permanent. It updated technical capabilities, which had stagnated for fifteen years, by developing three new systems, the Archival Electronic Records Inspection and Control (AERIC) for accessioning, the Archival Preservation System for preservation, and the GAPS database for managing transfers of electronic record to the National Archives. AERIC and APS introduced automation, bulk processing, and automatic generation and capture of management data and metadata into what had been very labor intensive, piecemeal processes, and provided greater flexibility in NARA's ability to provide access to electronic records. GAPS translated the essentially free-form textual information in records disposition schedules for permanent electronic records into structured data and used that data to project when records should be transferred to the National Archives. Initial analysis of this data produced two startling discoveries. The first was that, in 1989, the National Archives had received only 2% of the permanently valuable electronic records that should have been transferred by then. The second was that approximately 40% of the schedules for permanently valuable electronic records could not be interpreted to determine when transfers should be received. The Center set about improving both statistics. In most cases, the impossibility of projecting transfers was due to conditional disposition instructions; that is, rather than stipulating a date or fixed term for transfer, the instruction stipulated that records should be transferred when some condition was satisfied. In many of these cases, the problem could be eliminated, without revising the schedule, simply by asking the records creator how frequently the condition was satisfied and using that frequency to project transfers.³

The combination of increasing the identification of permanently valuable electronic records by appraisal and targeting records that should have been transferred in the past enabled the Center to substantially increase the National Archives' holdings of electronic records, as shown in figure 1.

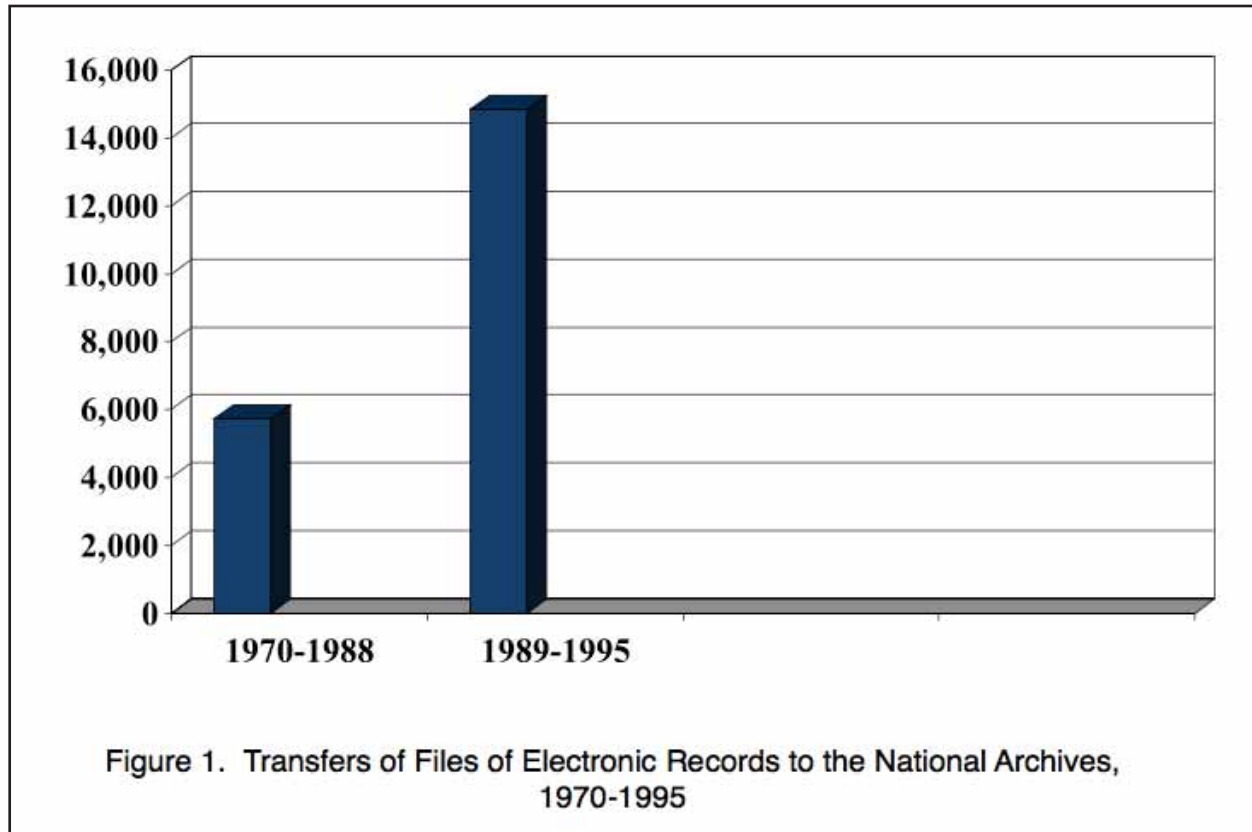
In the first two decades of dealing with electronic records, the National Archives had accessioned less than 6,000 files. In years, the Center for Electronic Records accessioned approximately 10,000 additional files, tripling the rate of transfers. The growth also reduced the gap between actual holdings and those projected by the GAPS database by a factor of ten. The new AERIC and APS systems enabled the Center to process the new transfers expeditiously.

The growth in holdings is even more impressive when one considers the shock wave that hit the Center in 1993. The shock came as the result of a lawsuit against the Executive Office of the President of the U.S., in which NARA was a co-defendant. Commonly referred to as "the PROFS case," the suit targeted email in the administrations of Presidents Reagan and George H. W. Bush on 6 January 1993, just two weeks before the end of President Bush's term of office, the court ruled that the digital versions of e-mail in federal agencies within the Executive Office of the President (EOP) met the statutory

² Kenneth Thibodeau, "Rupture ou continuité: l'évaluation des archives au seuil de l'époque numérique," *Archives* 31, no. 3 (1999-2000): 62-72.

³ Bruce Ambacher, "The Evolution of Processing Procedures for Electronic Records," in *Autorità per l'informatica nella Pubblica Amministrazione. La conservazione dei documenti informatici - Aspetti organizzativi e tecnici*. Seminario di studio Roma 30 ottobre 2000, pp. 7-14.

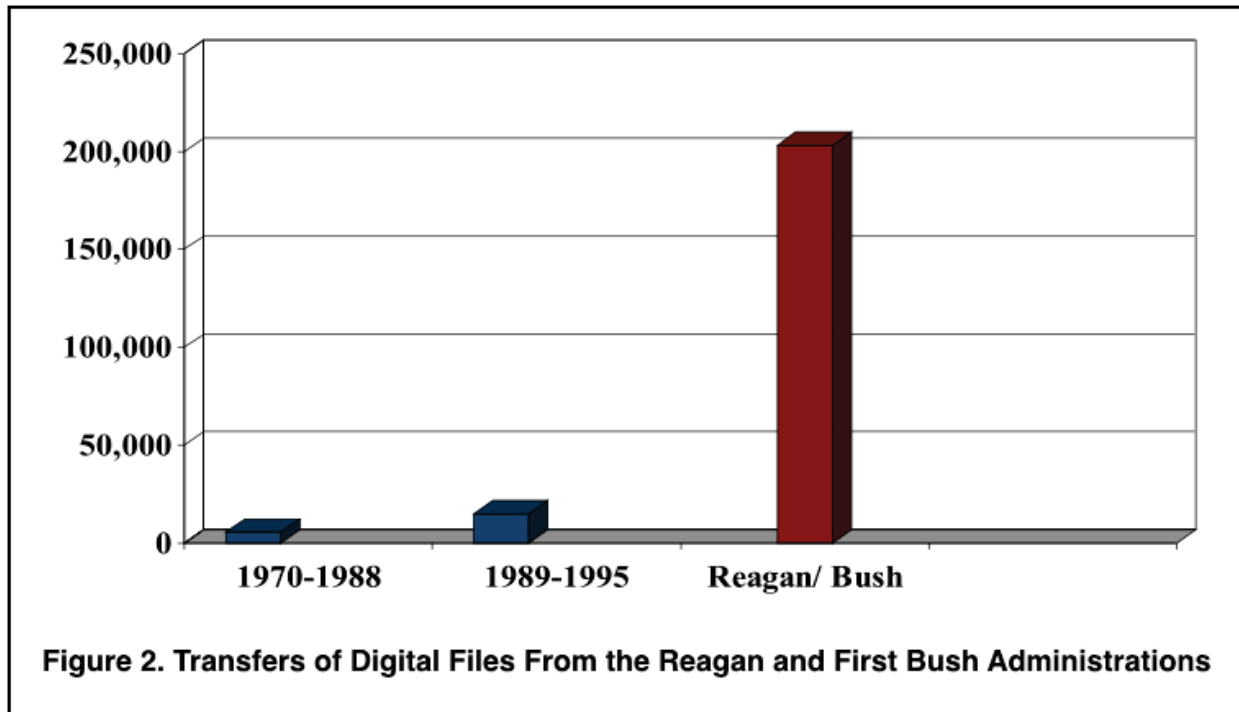
http://www2.cnipa.gov.it/site/_contentfiles/00309200/309235_documentazione.pdf; Theodore J. Hull, "Reference services and electronic records: The impact of changing methods of communication and access," *Reference Services Review* 23, no. 2 (1995): 73-78.



definition of ‘federal record.’ Over the next week the court issued specific orders requiring preservation of these electronic records.⁴ Up to that point, the EOP practice had been to print copies of email records and integrate them into paper filing systems. The court’s decision led to the transfer to NARA of over 200,000 digital files from the administrations of Presidents Reagan and Bush. Quantitatively, this dwarfed all of the electronic records that NARA had received since 1970, as shown in figure 2.

Moreover, qualitative difficulties entailed by this transfer greatly compounded those posed by the overwhelming quantity of files. First of all, the transfers came from a variety of computers ranging from PCs to mainframes, on a wide variety of media, including magnetic tape reels in different formats, a variety of tape cartridges and cassettes, hard drives that have been taken out of PCs and even disc packs. NARA had no experience with and no equipment for most of these media. Secondly, the transfer did not consist only of records. Rather they included every type of file that one might find on live systems and backup tapes. They included not only wide varieties of user created files, but also software and other system files, tutorials, help files, and even computer games. Sifting through all these files to identify and extract those that might be records was an immense task. A subsidiary, but substantial burden was to identify and eliminate duplicate copies of records. The vast majority of transferred media consisted of backup tapes from mainframe and mini computer systems. Records that users retained over time were repeatedly backed up, producing numerous copies. Additionally, many emails and other files were shared among staff, each of whom kept a copy. This difficulty was further compounded by uncertainty about the record status of files created or received by White House employees and contractors. Such files could be

⁴ Jason R. Baron, “The PROFS Decade: NARA, E-Mail and the Courts,” in Ambacher, op cit., pp. 105-137.



records, but they might also be personal materials, or they might be files related to a person's political, rather than governmental activity. Moreover, practically none of the files that might legally qualify as records had been subjected to any records management regime. Their classification had not been assigned and their disposition had not been determined.

The third major difficulty was that none of the current computer capabilities of the Center for Electronic Records or any other organization in NARA could be used for these files. The AERIC and APS systems were still in development and therefore not available. Current archival processing of electronic records was done at a computer service bureau, but several factors dictated against using the service bureau for the White House files. Most obvious was the fact that the files were subject to litigation. In fact, before NARA transferred responsibility for these files to the Center, the court found the agency in contempt for its handling of the materials. While the contempt order was overturned on appeal, every step taken by the Center was under close scrutiny by the judge and plaintiffs. A second reason not to use the service bureau was that, even apart from litigation, White House materials include sensitive information. NARA needed to minimize the risk of inappropriate disclosure of any sensitive information. Another consideration was that NARA had no information about the physical integrity of the media that had been transferred. Shipping the media back and forth to the service bureau would entail an unacceptable risk of irreparable damage to the fragile media.

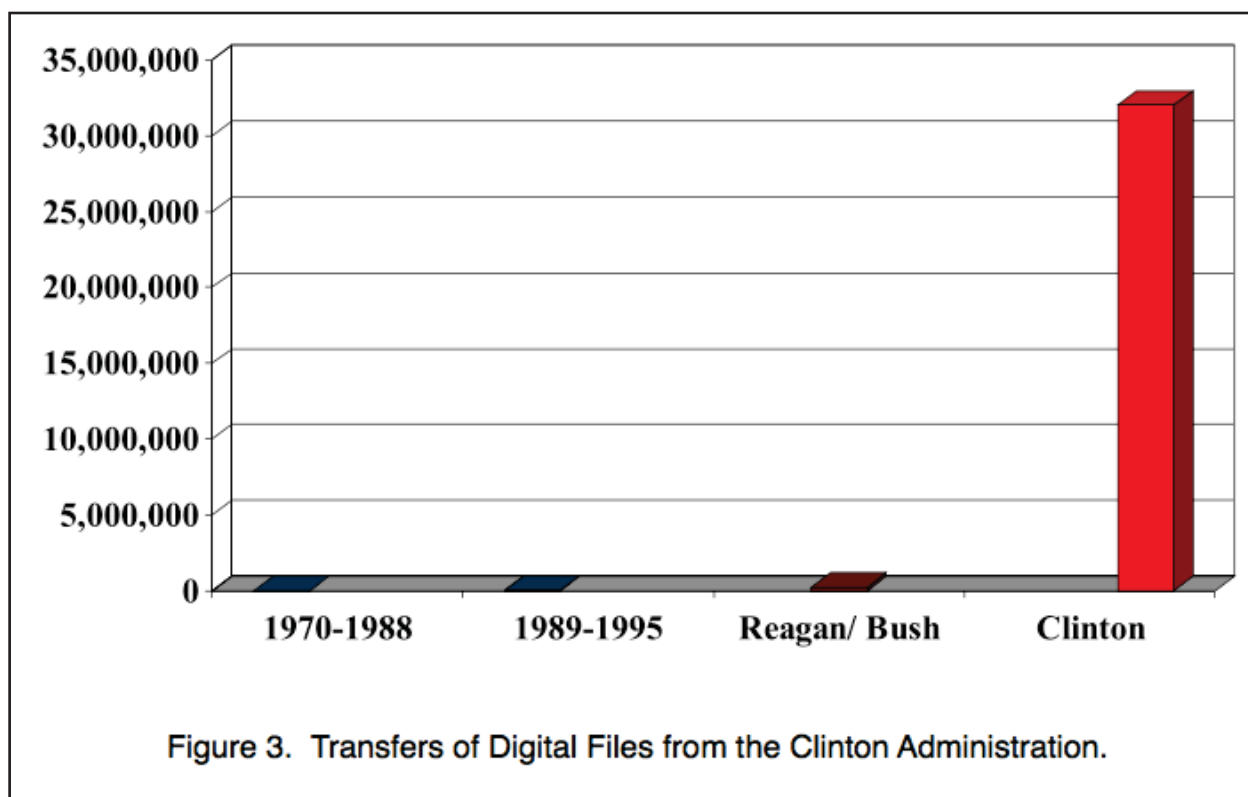
The fourth major difficulty was that the materials were subject to three distinct legal regimes: the Federal Records Act, the Presidential Records Act, and laws and regulations related to information classified for national security. Until then, NARA's activities related to electronic records had been concentrated almost entirely in the National Archives, which is responsible for federal records. NARA's archival responsibilities for presidential records are discharged by the Presidential Libraries. Due primarily to differences between the two records laws, NARA's business processes for federal and presidential records were very different. Moreover, national security concerns were a particularly vexing

problem because, in addition to having no equipment or software for processing the materials, the Center had no employees who had the clearances necessary to access this information.

Given this complex and burdensome situation, the Center for Electronic Records focused its efforts on two basic tasks: (1) ensuring the survival of all of the transferred files and (2) identifying and extracting all the files that might be records and making them available for additional archival processing. Ensuring survival entailed making copies of the files onto trustworthy digital media. Copying digital files is basically a simple task; however, the absence of any appropriate technology for performing the task coupled with the lack of essential information, such as the condition of the media, how data was physically inscribed on them, the formats of the files, etc., the task appeared extremely daunting, if not impossible. To attack the challenge, the Center asked the contractor responsible for developing the APS system, Mueller Media Conversions, Inc., to deliver as soon as possible a configuration of equipment for copying files from magnetic tape reels to the tape cartridges NARA used for preservation. The majority of the transferred files were on reels of tape. While this request was outside the scope of the contract, Mueller Media rose to the challenge and delivered a suitable configuration within a few weeks. NARA decided to start copying those files which had been the principal objective of the litigation in the PROFS case; namely, files from the National Security Council. Given that none of the Center employees had the clearances needed to access this information, the Center was authorized to copy the files only on condition that there was no possibility that anyone performing the work could access the content. Thus, the copying system was configured without any printer and in such a way that file content could not be displayed on the screen. This made it very difficult to deal with problems encountered in copying. There were many problems. The initial procedure, then, was to abort any copy job when a problem was encountered and proceed to copy the next reel of tape. Eventually, members of the Center's staff received the clearances necessary to access the information and the system was reconfigured to permit access to the data so that problems could be analysed and solved. Furthermore additional instances of the system were acquired from Mueller Media to deal with unclassified media and to increase the rate of production of copies. The systems was also modified incrementally to deal with additional physical media and additional physical and logical formats. Through such efforts the Center was able to copy successfully more than 99.9% of all the data on the tapes, and in the process to identify all user-created files. This success was a critical factor leading the court to dismiss the lawsuit in the government's favor.

Regardless of this success, NARA's experience in the PROFS case was so traumatic that the agency determined to do everything it could to avoid anything similar in the future. To this end, NARA staff were assigned to train every employee in the Executive Office of the President on the basics of managing records, with special emphasis on electronic records. In addition, from the beginning NARA worked closely with officials of the Clinton Administration in order to prepare for the eventual transfer of Clinton electronic records. In spite of these efforts, it became clear that NARA would face a tsunami when these records were transferred. Through its collaboration with the White House, by the midpoint of the Clinton Administration, NARA was able to project that the volume of electronic records to be transferred would make everything NARA had handled to that point look small. As illustrated in Figure 3, that proved to be the case. The Clinton materials were two orders of magnitude greater than all the digital files NARA had acquired previously.

The midpoint projection of Clinton transfers led NARA to a milestone decision. Even though the AERIC and APS systems had been developed, implemented and had enabled massive increases in productivity by that time, analysis showed that existing capabilities could not be scaled up to complete even the most basic task related to transfer of the Clinton records, the production of preservation copies.



A new approach was needed. Finding one would not be easy because many archival requirements related to preservation and sustained access to electronic records were beyond the state of the art of information technology. Some were even beyond the state of the art of computer science.

Thus, in August 1998, John Carlin, the Archivist of the United States, authorized the Electronic Records Archives Project, charged with conducting research to find ways to meet the ever expanding and increasingly complex challenges posed by electronic records.

The search for a solution had to look beyond the areas of archives and records management where NARA traditionally operated. The exploration began with a survey to identify automated systems in other federal agencies which, regardless of the purposes they served, had characteristics that could be adopted or adapted to meet NARA's needs. This search proved futile. NARA then turned its attention to the arena of high performance computing research. This venue proved more fruitful. The ERA project established long lasting collaborations with several major players in computer science and technology research in the U.S. government, including the Advanced Research Projects Agency of the Department of Defense (DoD), the National Science Foundation (NSF), the U.S. Army Research Laboratory, and the National Aeronautics and Space Administration. NARA also became one of the principal supporters of the International Research on Preservation of Authentic Records in Electronic Systems (InterPARES) project.⁵

⁵ Kenneth Thibodeau, "Preserving Digital Memory at the National Archives and Records Administration of the U.S." (presented at Workshop on Conservation of Digital Memories. Second National Conference on Archives, Bologna, Italy. 20 November 2009).

The collaborative approach taken in ERA research was reflected the strong commitment in NARA's Strategic Plan of 2000 to address records management challenges in the domain of electronic records in partnership with others in the federal government, state and local governments, and also in academe and the private sector.⁶ Through its collaborations, NARA sought not to develop technologies specific to archives, but to find solutions to archival problems in technologies that could serve as broad a range of interests as possible. These technologies would provide a framework in which an information management architecture for persistent archives could be developed. That architecture specified the framework to address archives and records management requirements, but was general enough to be applicable in other archival institutions besides NARA.⁷ Results of the collaborations were promising enough that, in January 2000, John Carlin raised the status of the Electronic Records Archives program from that of a project to that of a strategic initiative:

We can look forward to building another new archives, this one constructed from computer and communications systems. And it will not be located in any one place - it will stretch across NARA's nationwide system.

We face this task with a level of comfort that would have seemed foolhardy a few years ago. The comfort comes in part from technological advancements themselves, but mainly it stems from the fact that we can draw on the resources and the expertise of partners - in government and in the private sector, around the country and around the world.... Thanks to these collaborations, we have been able to lay out our vision for the new archives in sufficient detail that we have given it a name, the Electronic Records Archives (ERA), and can now set it in motion.⁸

For fiscal year 2002 Carlin requested and received a substantial increase in funding for the program in 2002. That funding and additional increases in the following years enabled NARA to build the ERA system. The ERA Program spent several years developing the capability to manage a system development that would far exceed anything NARA had ever done in the IT arena. In the same period, the program engaged representatives from the entire agency, as well as outside stakeholders, including other government organizations, researchers, the IT industry, and the public at large, in an iterative process of defining the requirements for ERA. In 2005 it conducted a competition between two coalitions of companies to articulate the system design. Development started in 2006, focusing on automating NARA's lifecycle management of federal records, both electronic and traditional. This laid the foundation for preservation and long-term access to electronic records. This system was put into operation in 2008.

In the interim, the program had started development of a second version of ERA specifically designed to process the huge volume of electronic records expected from the White House at the end of the administration of President George W. Bush. As stated above, presidential records are subject to different legal requirements than federal records. The regime of records disposition schedules, which results in a continuing stream of transfers of permanently valuable federal records to the National

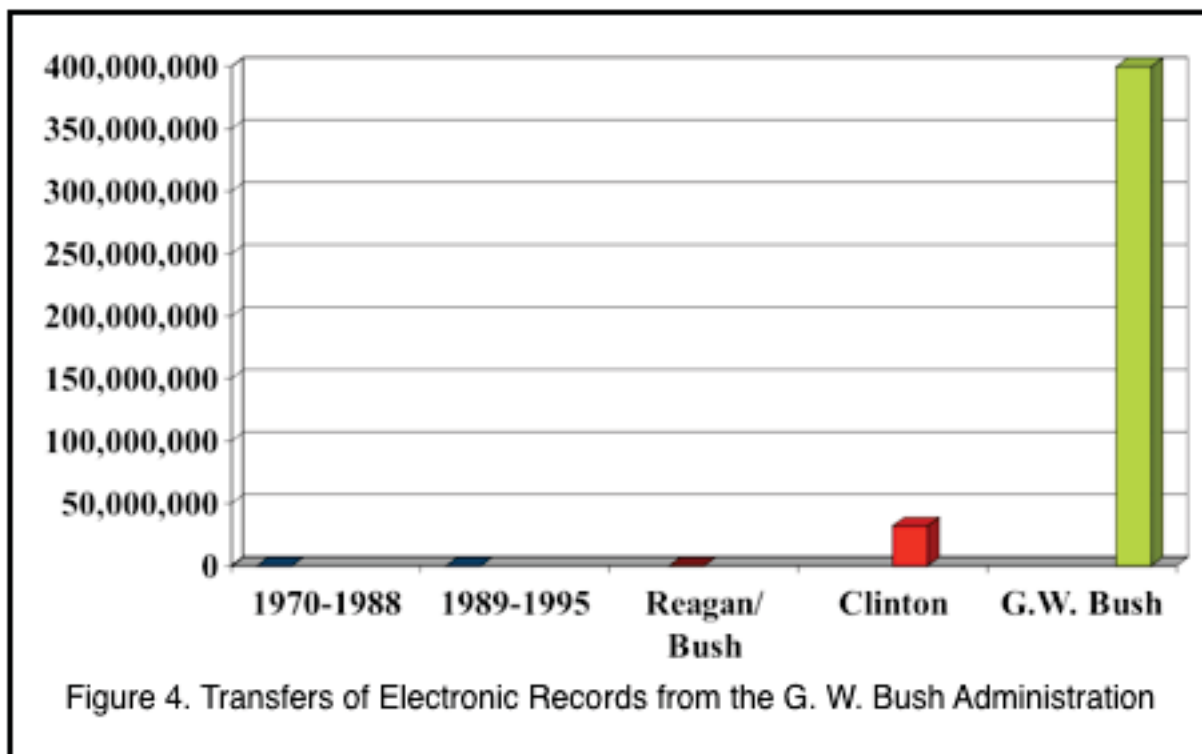
⁶ National Archives and Records Administration. "Ready Access to Essential Evidence: The Strategic Plan of the National Archives and Records Administration, 1997-2007" (Revised 2000). Available at: http://www.archives.gov/about_us/strategic_planning_and_reporting/2000_strategic_plan.html.

⁷ Kenneth Thibodeau, "Building the Archives of the Future: Advances in Preserving Electronic Records at the National Archives and Records Administration," *D-Lib Magazine* 7, no. 2 (February 2001). <http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html>.

⁸ NARA Notice 2002-74. Electronic Records Archives (ERA) Program. January 19, 2000.

Archives, is absent in the case of presidential records. Rather, when a president leaves office, all presidential records of that administration immediately become the legal responsibility of the Archivist of the United States. Furthermore, while the basic rule under the Freedom of Information Act is that federal records are open to the public, public access to presidential records is restricted for several years. But the current president, the former president, the Congress, and the courts have immediate rights of access, which they quickly and frequently exercise. These are important customers whose demands need to be met. However, the situation is complicated by the fact that, given the separation of powers embedded in the U.S. Constitution, access by the Congress and the courts is subject to conditions involving case by case approval by the current and/or former presidents. The second ERA system was designed to meet these requirements, with special emphasis on providing thorough, accurate, and prompt responses to such special access requests, while supporting content review to ensure that no information would be inappropriately disclosed.

Satisfying these requirements was complicated by the fact that once again NARA was confronting the perfect archival storm: an exponentially increased volume of records, including numerous formats with which NARA had never dealt, and a dearth of basic information needed to plan for processing and access. As illustrated in figure 4, once again NARA faced a transfer of electronic records that dwarfed everything that NARA had received in its history.



The transfers totaled over 72 terabytes of data in almost 400,000,000 files. They included more than 200,000,000 emails, more than 11,000,000 digital photographs, 48,000 digital motion videos, more than 29 million records of entry by workers and visitors to the White House complex, records from 36 other computer systems or applications in the EOP, as well as vice presidential electronic records and classified electronic records.

Given the requirements for the Bush records, what might have been a relatively simple task of moving the records from the EOP to NARA became a serious challenge. In the past, all electronic records, whether federal or presidential, had been transferred to NARA on physical media, but moving hundreds of millions of objects on media would not satisfy the requirements for prompt access to records; moreover, tracking all of those objects throughout the transfer would have been a labor intensive logistical nightmare. An alternative might have been to transfer the files over the Internet, but on analysis it was determined that there was not sufficient bandwidth between the White House computer center in Washington and the ERA center in West Virginia to complete the transfer expeditiously. Furthermore, security experts identified significant risks in this alternative. Engineers from the Lockheed Martin team responsible for developing and operating the ERA systems suggested a third possibility: putting ERA servers in the White House computer center, connecting them to the system bus, and copying the records directly from the EOP system to the ERA devices. Once a given server was full to capacity, it would be transported to the ERA computer center and attached to the ERA system, making the records instantaneously available for archival processing. NARA managers were initially dubious whether EOP officials would allow such access to their system, but the EOP proved receptive and quite cooperative, even upgrading the capacity of their system bus to expedite copying. They even allowed NARA to begin the transfer process a month before Inauguration Day, which is the date stipulated for transfer in the Presidential Records Act.

The ERA system for presidential records was built on the same architectural model as the initial ERA system, but the specific design and functionality were substantially different. The initial system was designed for lifecycle management of both hard copy and electronic federal records. It provided automated support for the processes of records scheduling and appraisal. All other functions supported by this instance were articulated on the assumption that all records in the system were under specified records disposition authorities. Given the Presidential Records Act's stipulation of the transfer of all extant presidential records at the end of each administration, records scheduling and appraisal were irrelevant in the EOP instance of ERA. Rather its design was driven by the need to respond quickly, thoroughly, accurately, and appropriately to requests from the presidents, Congress and the courts. The principal processes supported by the EOP instance are rapid ingest of large volumes of electronic records, automatic indexing on ingest, organization of records into ad hoc groupings defined to facilitate responses to anticipated requests, production of different versions of records to facilitate faceted search appropriate to different types of records, immediate ability to search, enforcement of access restrictions, case management for review and redaction of sensitive content, and detailed audit trails to ensure accountability. An additional, but critical capability of the system was the ability to isolate and remove any classified information discovered in the unclassified instance. Without this capability, it would have been necessary to bring down the entire system when any such discovery was made.

The EOP instance of ERA enabled NARA not only to weather the perfect archival storm, but also to achieve astounding success in meeting its requirements for the G.W. Bush electronic records. Between December 12, 2008 and September 30, 2009, NARA ingested more than 267 million objects into the EOP ERA. Ingest included scanning the objects for malware or other technical problems, creating full text indexes, and organizing them into access groups. During this processing, ERA identified 65 million problems of various sorts. For example, viruses were detected in over 1,250,000 files; thousands of files had no content; 36,000 email messages were corrupted; in the collection of digital photographs, there were so many missing images that the White House had to redo the entire transfer; and the records management system, which is the finding aid to White House paper records, was repeatedly transferred in unusable formats. Working collaboratively during the 10 month ingest process, NARA and White House

staff were able to eliminate more than 99% of all these problems. The majority of unresolved problems were cases where viruses or other malware were detected by ERA. Based on a manual sampling, security experts concluded that most probably these were cases where ERA had detected residues of malware that had actually been removed previously by the White House system. Rather than risk infecting the ERA system, all of these case were exported for offline processing. The success of the EOP ERA was not limited to ingest and detection and resolution of problems. During ingest, ERA created almost 230,000,000 new versions of records to facilitate faceted search. Almost 100,000,000 duplicate records were identified and eliminated, all under precise audit trail. By September 2012, the 26 NARA staff who work with the Bush records had executed 66,000 searches in the system.⁹

Given the predicted archival storm, presidential records staff were appropriately fixated on the transfer of the Bush records; however, once the ERA system had proven its ability to meet the immense challenges of this transfer, responsible officials began plans to extend its capabilities to other presidential libraries with substantial collections of electronic records, most notably to the library that will be built for the records of President Obama. Nevertheless, while ERA is seen as a game changer for NARA, the agency will undoubtedly face other traumatic experiences with electronic records in the future. Given the finite size of the Executive Office of the President, the growth in the volume of email is likely to slow down and perhaps even reach a limit. However, it is equally likely that the White House will continue to rely increasingly on digital information and communications technologies. Thus it will continue to generate more and more electronic records and it must be expected that many of these records will be in new formats that had not been transferred to NARA previously. There is already a sign of the archival storm that will accompany the next transfer of presidential electronic records to NARA: the Obama Administration's unprecedented use of social media.

⁹ About the Executive Office of the President Instance (EOP). <http://www.archives.gov/era/about/exec-office-instance.html>

Trusting Records

The Ethical and Legal Issues of Historical Mental Health Records as Cultural Heritage

Lorraine Dong

Abstract

This paper will examine the ongoing repercussions of United States health privacy laws on the retention of historical medical records and their integration into our cultural heritage landscape. The paper will begin with a broad discussion of whether records from medical facilities should be considered heritage belonging to everyone. Questions of legal, ethical, and cultural ownership and privacy will be addressed. The author will then present Central State Hospital, a 142-year-old mental institution in Petersburg, Virginia, as a case study of an institution and its historical records in transition. The author is part of a team from the University of Texas at Austin working to preserve and provide access to the records through both public and restricted digital interfaces. She will conclude with matrices of possible archival actions that address the limitations and unexpected potentials of working with such documents, such as the inclusion of emergent voices in the archival process.

Author

Lorraine Dong is a doctoral candidate in preservation and archives at the University of Texas at Austin's School of Information (UT). She received a BA in English literature from the University of California, Berkeley, an MPhil in Renaissance literature from the University of Cambridge, and a MSIS from UT. Lorraine has published articles in *Archival Science* and *Archival Issues*.

1. Introduction

Whether historical mental health records should be preserved and made accessible as part of our cultural heritage should depend on the situational context. The current reality, however, is that sweeping privacy laws and the holding institutions' policies in response to those laws determine the fate of these documents. At present in the United States, mental health documents remain in a records category that is generally off-limits to the public, researchers, and family members. The question of who should be granted access to the materials remains an especially problematic and fundamental issue for any preservation and access initiative for records with sensitive information.¹ It is the hope of this author that an examination of these records as heritage will encourage a more nuanced discussion about the care, control, and ownership of any archival collection containing restricted documents. This paper will present the legal and cultural landscape in the U.S. for the management of medical information. Then the author will discuss the Central State Hospital archives project as a case study on the ethical considerations heritage professionals must give to the many communities tied to the records. The conclusion will focus on potential solutions made possible by digital technologies and international archival scholarship.

¹ Archivists are just beginning to produce articles regarding this subject. See Anne T. Gilliland and Judith A. Wiener, "Digitizing and Providing Access to Privacy-Sensitive Historical Medical Resources: A Legal and Ethical Overview," *Journal of Electronic Resources in Medical Libraries* 8, no. 4 (2011): 382-403.

2. Medical Records and Privacy in the United States

In the U.S., institutional records that specifically pertain to individuals who are vulnerable and/or contain informational categories deemed private by the federal and state governments remain legally shielded from public view without much nuance. Examples of such records include student files that fall under the Family Educational Rights and Privacy Act of 1974 (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99), closed adoption papers, and health records. These records may contain information such as social security numbers, billing details, and names of those who wish to remain anonymous. Often perceived as personal and private, such records contain information considered to be viewable only by a select number of people, such as the patient, immediate family, and medical staff.

The protection of personally identifiable information in general became a growing concern in the 1960s during the rapid growth of computer use by public and private organizations to collect and manage personal data; acts such as FERPA and the Privacy Act of 1974 were ill prepared to respond to technologies that can locate individuals through large data sets.² The Health Insurance Portability and Accountability Act of 1996 (PL 104-191), or HIPAA, and its regulation, the Privacy Rule of 2002 (45 CFR §160, §164), were ostensibly created to protect the security of health information in increasingly digital healthcare environments that now include third-party companies and multiple providers. As a consequence of HIPAA and the Privacy Rule, any person, business, or agency that provides, bills for, or receives medical care payment is responsible for protecting private health information. All medical facilities that use electronic health records at the time HIPAA was passed are considered to be “covered entities,” while hospitals that were closed prior to the passing of HIPAA do not need to adhere to the law. HIPAA and the Privacy Rule dictate state-level laws regarding mental health privacy, with some states such as Virginia being very strict in its medical privacy laws, while others such as Texas are more lenient.

Unlike any past major regulation in the U.S., the Privacy Rule applies to all records, regardless of when they were created, into perpetuity. Thus, if a record under the jurisdiction of a covered or hybrid entity (e.g., medical universities that have areas falling under both covered and non-covered jurisdictions) contains what is considered private health information, that information must be restricted, regardless if it pertains to an individual who died five years or 150 years ago. HIPAA and its regulations do not provide much leeway for researchers, including patients’ families and descendants, to gain access to the records. The only exceptions for attaining access to these records are if the restricted information has been redacted, information has been aggregated into large datasets thus making the identification of individuals extremely difficult or impossible, or the researcher has received a waiver of authorization from a Privacy Board.

Although federal and state health privacy laws in the U.S. attempt to have unambiguous delineations, they are couched in complex cultural and political frameworks of individualism and privacy. It is undeniable that the data sets that could be produced from medical records would “play an important role in research, health care, data security, and the dissemination of knowledge generally.”³ Less clear is the cultural value of revealing the identities of institutionalized patients and their records, especially when weighed against the ethical responsibility to respect the individual and her family’s right to privacy. In particular, while most individuals would understandably want to keep their medical records relatively private such that only a limited group of people are allowed access to them, the answer is certainly

² Paul M. Schwartz and Daniel J. Solove, “The PII Problem: Privacy and a New Concept of Personally Identifiable Information,” *New York University Law Review* 86 (2011): 1814-1894, pp. 1820, 1824.

³ *Ibid.*, 1866.

blurrier when considering health records of individuals who have died many decades ago. It should be remembered that notions of privacy are, like heritage, both culturally and temporally fluid. While some communities may view mental illness as an extremely personal matter that should be dealt with (and sometimes kept secret by) close family members, others may openly recognize it and want to share their findings.⁴ Furthermore, what may have been considered private information a century ago is now considered common knowledge, and vice-versa.

At present, politicians and others who determine the laws for the preservation of and access to inactive medical records put a greater value on individual privacy rights, especially those in relation to potentially stigmatizing information, over the interests of researchers, ancestors, and the public in general. Currently, the U.S. federal government is considering strengthening the protections for research that may “pose informational risks” by adopting HIPAA-level restrictions for information in non-covered entities.⁵ Individual privacy and the intimate relationship between doctor and patient are privileged over familial relationships and public knowledge. Daniel Solove contends that privacy is not an individualistic right; rather, “constitutive privacy understands privacy harms as extending beyond the ‘mental pain and distress’ caused to particular individuals; privacy harms affect the nature of society and impede individual activities that contribute to the greater social good.”⁶ Solove posits that identification can negatively impact identity by attaching “information baggage to people” and ending anonymity; he emphasizes that it is not just the recorded individual at risk but all people associated with that individual.⁷ Based on Solove’s emphasis on the shared aspect of privacy, it is arguable that those who are not patients but nevertheless potentially affected by an individual’s health records should have a voice in how those records are handled, especially if the individual is deceased. Ironically, HIPAA acts as a deterrent for research that would examine the attitudes of families and descendants of deceased patients toward the possibility of having their ancestors’ records released, either to the public or to select individuals and groups.

3. The Central State Hospital Archives Project

The balancing of privacy rights of patients and their families with the recognition of medical records as documents of potentially great historical and cultural value is a critical concern to the archival project that serves as this paper’s case study. The primary site of the author’s dissertation research is Central State Hospital (CSH), a state mental institution in Petersburg, Virginia. The hospital is most notable for its role in African American mental health care in the U.S. It was the first hospital specifically for Black patients when it was created in 1870 by the Virginia Legislature under pressure from the Freedman’s Bureau. Originally called the Central State Lunatic Asylum for the Colored Insane, the hospital’s identity has been and continues to be shaped by its African American origins, whether openly acknowledged or not. While

⁴ A very public example of an individual learning about an institutionalized ancestor and attempting to pursue more information is the actor Blair Underwood, who participated in a 2012 episode of the NBC television show, *Who Do You Think You Are?*, and discovered that his three-times great grandfather, Sauney Early, was a patient at Central State Hospital in Petersburg, VA, during the turn of the century.

⁵ Department of Health and Human Services, “Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators (HHS-OPHS-2011-0005),” *Federal Register* 76(143) (2011): 44, 521-44, 531, p.526.

⁶ Daniel Solove, “A Taxonomy of Privacy,” *University of Pennsylvania Law Review* 154, no. 3 (2006): 477-564, p. 488.

⁷ *Ibid.*, p. 513.

some scholars have stated that mental institutions in the U.S. and Europe were usually located far from towns because of the stigma of contagion and difference attached to mental illnesses,⁸ the relocation in 1885 of the hospital from Howard's Grove near Richmond to just outside of Petersburg was seen as an economic and social boon by Black and White leaders of Petersburg. According to an unpublished history of CSH edited in 1960 by Dr. Theodore Denton, a former superintendent of the hospital, the city of Petersburg offered to purchase and donate the land to the state of Virginia so a new mental institution could be built close to the town.

CSH began with a few hundred patients. Patient ranged from the elderly and indigent to those with tuberculosis or with mental illness. Some were placed at the hospital simply because they were perceived as lazy or defiant. The commitment process involved a short questionnaire to be answered by an authority figure, neighbor, or family member, and served as legal evidence of an individual's insanity. Using a treatment method popularized by Philippe Pinel's work in the late 18th century, patients were not physically restrained but instead given "moral treatment" that called for the separation of individuals from their usual communities in order to impress them with "normal" behaviors.⁹ Thus, early patients at CSH worked for the hospital, albeit without pay, and conveniently provided a source of income for the institution. Photographs, annual reports, and board minutes provide evidence of patients growing crops, raising animals, working in the kitchen and storage warehouses, sewing, and otherwise contributing hospital labor. Patients were given a relatively large amount of freedom to roam the large hospital campus, resulting in many recorded escapes. At the turn of the century, the hospital began separating patients into wards that went beyond the male/female distinction, and an effort was made to distinguish the mentally ill from the mentally deficient in order to provide different treatments.¹⁰ The hospital continued to grow rapidly to accommodate an ageing population, and swelled to over 5,000 patients in the 1950s. Due to overcrowding, the focus for nurses shifted from mental health care to simply providing sufficient physical care, e.g., bathing, turning patients.

Several years after the Civil Rights act in 1964, the hospital was desegregated. While the hospital always had both African American and Caucasian nurses, staff, and volunteers to varying degrees (African American doctors and administrators began to be hired in the 1970s), the patients were exclusively classified as African American or mixed race. By 1968, White patients from other Virginia state hospitals were being sent to CSH. While race presumably no longer mattered when it came to the location of treatment, and by inference, the quality of care, recordkeeping categories at CSH indicate that race persisted as an important characteristic for administration. Today, the hospital continues to serve the central Virginia area. CSH primarily treats forensic patients, who have been either deemed by the court to be unfit for trial or ordered to serve time at the hospital. In addition to the several hundred forensic patients, some mentally deficient patients continue to live on the grounds in special assisted living cottages. Rapid changes in funding allocations will most likely lead to the hospital phasing into a correctional facility within the next decade.

In the winter of 2009-2010, federal grant money allowed a multidisciplinary team from the University of Texas at Austin to begin a CSH records preservation and archives project. The author joined the project

⁸ Gerald Grob, *Mental Illness and American Society, 1875-1940* (Princeton: Princeton University Press, 1983), 166; Michel Foucault, *History of Madness* (London: Routledge, 2006), 86.

⁹ Gerald Grob, *Mental Institutions in America: Social Policy to 1875* (New York: Free Press, 1972), 42.

¹⁰ Steven Noll, "Southern Strategies for Handling the Black Feeble-Minded: From Social Control to Profound Indifference," *Journal of Policy History* 3, no. 2 (1991): 130-151, p. 136.

in the spring of 2010 as the archivist who would process the existing collection at the hospital and prepare the materials for digitization. She continues to be involved in the development of the digital database and archives, and the necessary discussions regarding the ethical and legal privacy issues of the collection.

The collection consists of a wide range of materials that served a number of functions for the hospital, including admission records, board minutes, ward books, and annual reports, beginning from the hospital's inception. There are also photographs, manuals of practice, financial documents, news clippings, staff newsletters, and guest registers. Nearly all of the existing early patient records are on microfilm rather than paper, as the latter were intentionally destroyed in the mid-20th century as a space-saving measure after being microfilmed. The dates for the collection materials range from 1870 to 2010, although the project's primary focus is on the hospital's first 100 years from the end of the Civil War to the advent of the Civil Rights Act and subsequent patient desegregation. The physical collection housed at the hospital is composed of over 70 linear feet of historical material.¹¹ Due to donations by staff and discoveries of archival materials throughout the extensive hospital campus, the collection will most likely continue to grow. There are over 4,112 unique items.

The CSH team has moved forward to safeguard the collection, despite many of the records' uncertain retention statuses, through physical preservation and conservation and the production of digital surrogates (Fig. 1). The latter will be placed in a dark archives or one limited to CSH staff use only. The goal of broader public access, however, is muddled with privacy concerns and questions of physical, intellectual, and cultural ownership. The team will have to negotiate the challenges of building a database and developing the information architecture and retrieval for a digital archives that can recognize both restricted information and privileged users.

At present, the physical collection has had very limited use. At the hospital, doctors occasionally use the admission registers as evidence of historical mental illness diagnoses when instructing nursing and medical students. Descendants of former patients intermittently seek out information about their relatives by contacting the hospital's records department, which can provide record copies depending on the specific legal circumstances of the inquiry. If the CSH collection was made public through physical and digital access, it would be a boon to researchers of African American history and medical historians, among others. However, as indicated by the first section of this paper, the legal repercussions and, perhaps more importantly, the social ramifications need to be examined in order for a workable solution to be implemented.

4. Matrices for Ownership, Preservation, and Access: A Discussion

Historical mental institution records are especially problematic archival materials because of their sensitive subject matter. In addition to considering the record creator, archivists and other heritage professionals must also be mindful of the record subject. Many of the records in this genre pertain specifically to individuals, i.e., mentally disabled or mentally ill patients, who were unable to make decisions about the handling of their active and inactive records. When the person has no agency, either because he is mentally unable to comprehend the impact of his own record information or because the record subject is deceased, the agency currently falls on hospital or archives administration in conjunction

¹¹ 52.4 cubic feet of CSH materials were donated to the state archives, the Library of Virginia, in 2005. The bulk of the documents are commitment papers from 1874 to 1906. Other materials include admission registers, deeds and leases, and meeting minutes. <http://ead.lib.virginia.edu/vivaxtf/view?docId=lva/vi00940.xml>

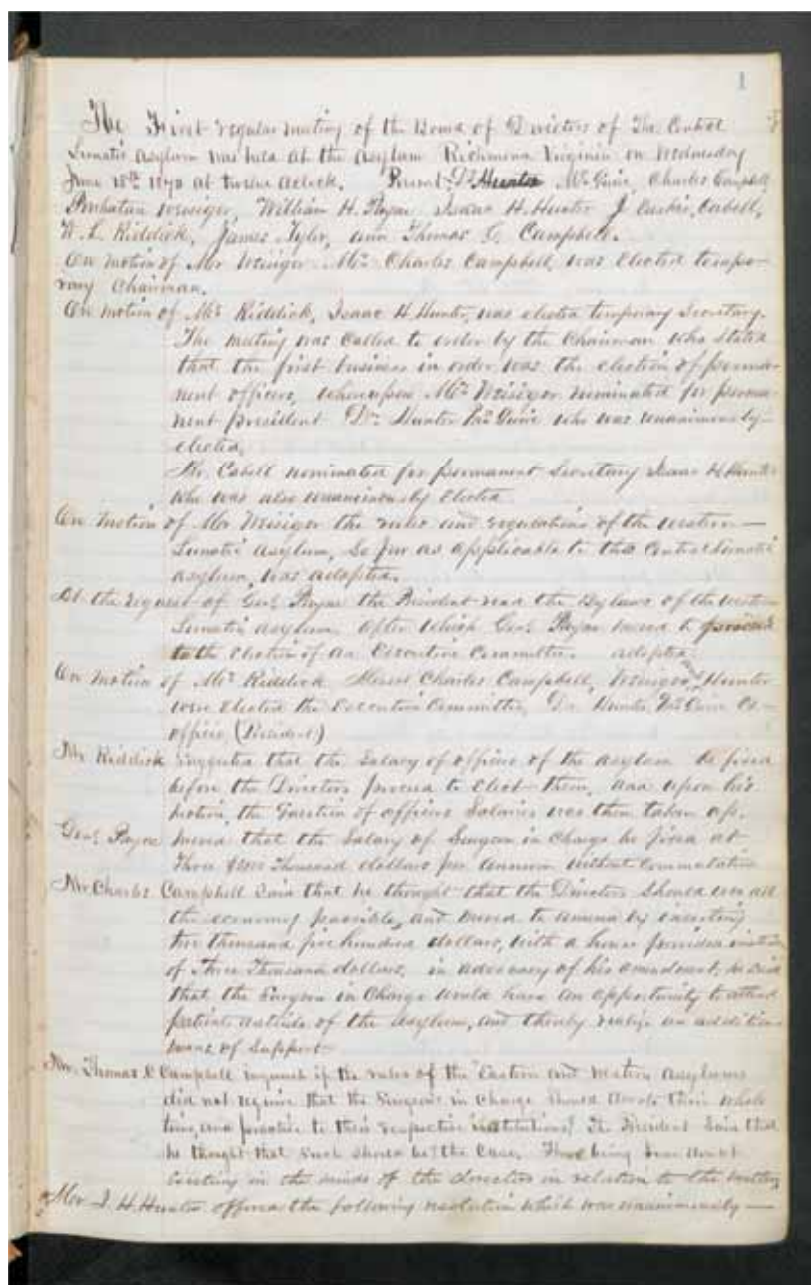


Figure 1. An example of CSH collection materials that have been digitized—the first page in a volume of hospital board meeting minutes, 1870 to 1884.

with institutional policies and governmental laws. In thinking about the archival ethics necessary to protect innocent parties who may be affected by archival record use, a number of archival scholars have asked, “To whom do the records belong?”¹²

¹² Terry Cook, “Professional Ethics and Practice in Archives and Records Management in a Human Rights Context,” *Journal of the Society of Archivists* 27, no. 1 (2006): 1-15, p. 10. See also Jeannette A. Bastian, *Owning Memory: How a Caribbean Community Lost Its Archives and Found Its History* (Westport: Libraries Unlimited, 2003); and Eric Ketelaar, “Sharing: Collected Memories in Communities of Records,” *Archives & Manuscripts* 33 (2005): 44-61.

In order to better understand the complexities of managing historical institutional records, the ownership of these materials should be considered from several perspectives. The term “ownership” can encompass multiple inclusive meanings, including the physical retention of the materials, legal control over the records or the information, and the cultural claim over the records as heritage artefacts. Depending on the context, records can arguably belong to one person, a community, a governmental institution, or a corporate entity. In a very broad sense, community is an interdependent social grouping bounded by a set of cultural characteristics and geography or resource space.¹³ What may be possessed physically by one group can be claimed culturally or intellectually by another.

The third type of ownership, the cultural right over tangible or intangible heritage, is especially imprecise. Similar to the idea of ownership, “cultural heritage” is a highly mutable concept that is dependent on societal norms and expectations, economic factors, and political objectives.¹⁴ What constitutes heritage depends on what is accepted by social groups as something that should be valued, shared within the groups, and passed on to future generations. In a symbiotic relationship, social groups delineate themselves through culture; in turn, culture is defined by its owners. The notion of cultural heritage as “property” that can be owned by an individual or a group continues to be a contentious topic among heritage professionals because of the implication that economic valuation can be placed on heritage.¹⁵ The commodification of heritage, while the basis of many much needed livelihoods, can also give rise to the exploitation of vulnerable social groups. For heritage record artefacts, the silent subjects of the materials must be considered in the ownership matrix.

While the necessity for protecting patient privacy and especially of those with little or no agency is perhaps obvious, the unyielding legal dichotomy of what information is considered “public” and what is “private” in the United States prevents most health information from becoming available to patient families, let alone researchers. Current records management and archival policies for most Northern and Western institutions do not reflect the contextual nuances of multiple types of ownership when addressing access.¹⁶ In Denmark, the Danish State Archives’ has an 80-year restriction rule on records that contain private and personal information, such as mental institution records.¹⁷ The Library of Virginia has a similar 75-year rolling restriction policy with its mental institution records holdings. Both of these access policies are perhaps necessarily straightforward, impersonal, and arbitrary for practicality’s sake.

¹³ Cf., Benedict Anderson, *Imagined Communities* (New York: Verso, 1991); John H. Freeman and Pino G. Audia, “Community Ecology and the Sociology of Organizations,” *Annual Review of Sociology* 32 (2006): 145-169, p. 145; and Anselm Strauss, Leonard Schatzman, Rue Bucher, Danuta Ehrlich, and Melvin Sabshin, *Psychiatric Ideologies and Institutions* (London: The Free Press of Glencoe, 1964), 38.

¹⁴ UNESCO, *Draft Medium-Term Plan (1990-1995)* (Paris: UNESCO, 1989), 57.

¹⁵ Mark Greene, “The Messy Business of Remembering: History, Memory, and Archives,” *Archival Issues* 28, no. 2 (2003-2004): 96-104; Heather Gill-Robinson, “Culture, Heritage and Commodification,” in *Cultural Heritages as Reflexive Traditions*, ed. Ullrich Kockel and Máiréad Nic Craith (New York: Palgrave MacMillan, 2007), 183-193; Alan Peacock and Ilde Rizzo, *The Heritage Game: Economics, Policy, and Practice* (New York: Oxford University Press, 2008); and Dallen J. Timothy and Gyan P. Nyaupane, *Cultural Heritage and Tourism in the Developing World: A Regional Perspective* (New York: Routledge, 2009).

¹⁶ Future works needs to be done to examine the constructions of “privacy” for mental health records in non-Western regions. Such an inquiry would have to reside within a larger examination of non-Western attitudes toward mental illness.

¹⁷ Inge Bundsgaard, “The Question of Access: The Right to Social Memory Versus the Right to Social Oblivion,” in *Archives, Documentation and Institutions of Social Memory*, ed. Francis X. Blouin, Jr. and William G. Rosenberg (Ann Arbor: University of Michigan Press, 2006), 118.

Despite the current legal landscape, the tantalizing potential exists for policies and technologies that would be minimally invasive on other archival tasks while maximally fulfilling an ethical principle that respects the social relationships among people and with records. The Society of American Archivists recently revised its Code of Ethics to reflect an increasing awareness among archivists of the contextual and cultural nature of privacy. While the code acknowledges that “privacy is sanctioned by law,” it also reminds archivists to be mindful of “individuals and groups who have no voice or role in collections’ creation, retention, or public use” and the privacy requests of “communities of origin.”¹⁸ Verne Harris, in discussing Derrida’s concept of “hospitality,” suggests an archival ethics that recognizes the constant need for re-contextualization based upon the particular relationships between the stakeholders and materials at any given moment.¹⁹ Erik Ketelaar provides the compelling possibility of a rights-based analysis conducted on a case by case basis in which the archivist bases her decision on a two-question human dignity test.²⁰ The first question is, “To what risk is human dignity exposed by public access to or publication of data which was provided confidentially?” and the second asks, “Is that risk acceptable - weighed against the expectation of any tangible or compelling social benefit, primarily to the records subjects, and secondarily to society as a whole?” Similarly, a contextualized approach toward information access can be seen in indigenous information management systems. For example, the “continuum of access” to the Mukurtu Wumpurrarni-kari Archive for the Warumungu Aboriginal community in Australia is based on a “dynamic system of accountability” that considers a member’s community status over his/her lifetime.²¹ Limitations to access can be the desired result of careful consideration toward the recorded community and the community of users.

The Mukurtu Wumpurrarni-kari Archive also points to the technological innovations that can serve as tools to address the complexities of access and ownership to privacy-restricted heritage materials. Isto Huvila highlights the possibility of having “a multitude of entry points to the information,” especially since the designated communities for archival participation are not defined until after a digital archives is made visible online.²² For indigenous digital archives and potentially with hospital archives, users’ interactions with the digital archives’ content depends upon each individual’s pre-existing relationship with the community portrayed in the records. In the vein of Barney Glaser and Anselm Strauss’ work on social interactions,²³ the users’ relationships with the online materials depend upon the “knowledge sets” they bring to the situation.²⁴ External cultural and legal restrictions that determine her status will pre-emptively decide what a particular user can access.

¹⁸ “SAA Core Values Statement and Code of Ethics,” Society of American Archivists, last modified January 2012, <http://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics>. See also Verne Harris, “Ethics and the Archive: ‘An Incessant Movement of Recontextualisation,’” in *Controlling the Past: Documenting Society and Institutions*, ed. Terry Cook (Chicago: Society of American Archivists, 2011), 349.

¹⁹ Verne Harris, *Archives and Justice: A South African Perspective* (Chicago: Society of American Archivists, 2007), 257.

²⁰ Eric Ketelaar, “Archives of the People, by the People, for the People,” *S.A. Archives Journal* 34 (1992): 5-16, p. 5, accessed August 18, 2012, *Academic Search Complete*, EBSCOhost.

²¹ Kimberly Christen, “Opening Archives: Respectful Repatriation,” *The American Archivist* 74, no. 1 (2011): 185-210, p. 189.

²² Isto Huvila, “Participatory Archive: Toward Decentralised Curation, Radical User Orientation and Broader Contextualisation of Records Management,” *Archival Science* 8, no. 1 (2008): 15-36, p. 18.

²³ Barney G. Glaser and Anselm L. Strauss, *Awareness of Dying* (Chicago: Aldine Publishing Company, 1965), 274.

²⁴ Christen, “Opening Archives: Respectful Repatriation,” 207.

5. Conclusion

As with all archives, the CSH collection can only provide “imperfect windows” into the past.²⁵ The CSH collection in particular is imperfect because of the large amount of privacy restricted materials. Until adjustments are made to HIPAA and state-level laws to better accommodate historical research in a contextualized manner, it is too risky ethically and legally for the hospital, the Virginia state archives, and the CSH project team to allow public access to records that focus on specific patients. However, this apparent lacking in access to many of the hospital-created documents can be viewed as an impetus for the active collecting of oral narratives and metadata. A current postmodernist trend in archival studies acknowledges the power of archives and the archivist, and calls for a proactive approach to building archives and archival communities.²⁶

Terry Eastwood suggests in his social theory of appraisal that contemporary usefulness is a collection value. Such usefulness goes beyond evidential value, and into the broad realm of public memory. Richard Cox elaborates the notion of contemporary use by saying that “archives are a symbolic way station on the road to a collective memory.”²⁷ In contrast to historical memory, Cox characterizes collective memory as constantly in transformation as individuals respond to intangible and tangible heritage. This selective process of creating oral and physical expressions of remembrance can be called “commemoration.”²⁸ Some recognition of CSH’s history already exists in centennial materials in the archival collection and physical structures on the hospital grounds (Fig. 2).

The development of commemorative value for the CSH collection is particularly important for the hospital as it transitions from its working hospital status to a historical site. The archival endeavor of building commemorative value has the potential to encourage collective memory across an array of communities and bring them in conversation with one another. The addition of non-hospital records to the collection of publically viewable institutional records will change the public value of the Central State archives. For instance, while the public website currently cannot serve a genealogical purpose, it is the author’s intent to provide contextualization of the records through the active collecting of alternative perspectives through oral narratives from long-time staff members of various departments and rankings. Furthermore, the creation of illustrative digital surrogates could stand as examples of the various forms used by the hospital without revealing restricted information. In this manner, the CSH website would be akin to a digital museum. The added value of oral narratives and other content and contextualization will give the collection greater focus on the hospital’s history, how it functioned as an organization, its place within the larger community, and how it is remembered as a historical place. Groups that are interested in building collective memory, whether they are the institutional administration, the archival custodians, or the local community, will ideally be able to use the CSH digital archives to construct their own commemorative works and contribute to the dialogue about the hospital’s role in history and today.

²⁵ Terry Eastwood, “Toward a Social Theory of Appraisal,” in *The Archival Imagination: Essays in Honour of Hugh Taylor*, ed. Barbara Craig (Ottawa: Association of Canadian Archivists, 1992), 83.

²⁶ Gerald Ham, “The Archival Edge,” *The American Archivist* 38, no. 1 (1975): 5-13; Terry Cook, “Archival Science and Postmodernism: New Formulations for Old Concepts,” *Archival Science* 1 (2001): 3-24; Huvila, “Participatory Archive”; and Katie Shilton and Ramesh Srinivasan, “Participatory Appraisal and Arrangement for Multicultural Archival Collections,” *Archivaria* 63 (2008): 87-101.

²⁷ Richard Cox, *No Innocent Deposits: Forming Archives by Rethinking Appraisal* (Lanham, MD: Scarecrow Press, 2003), 234.

²⁸ Jeannette A. Bastian, *Owning Memory*, 54.

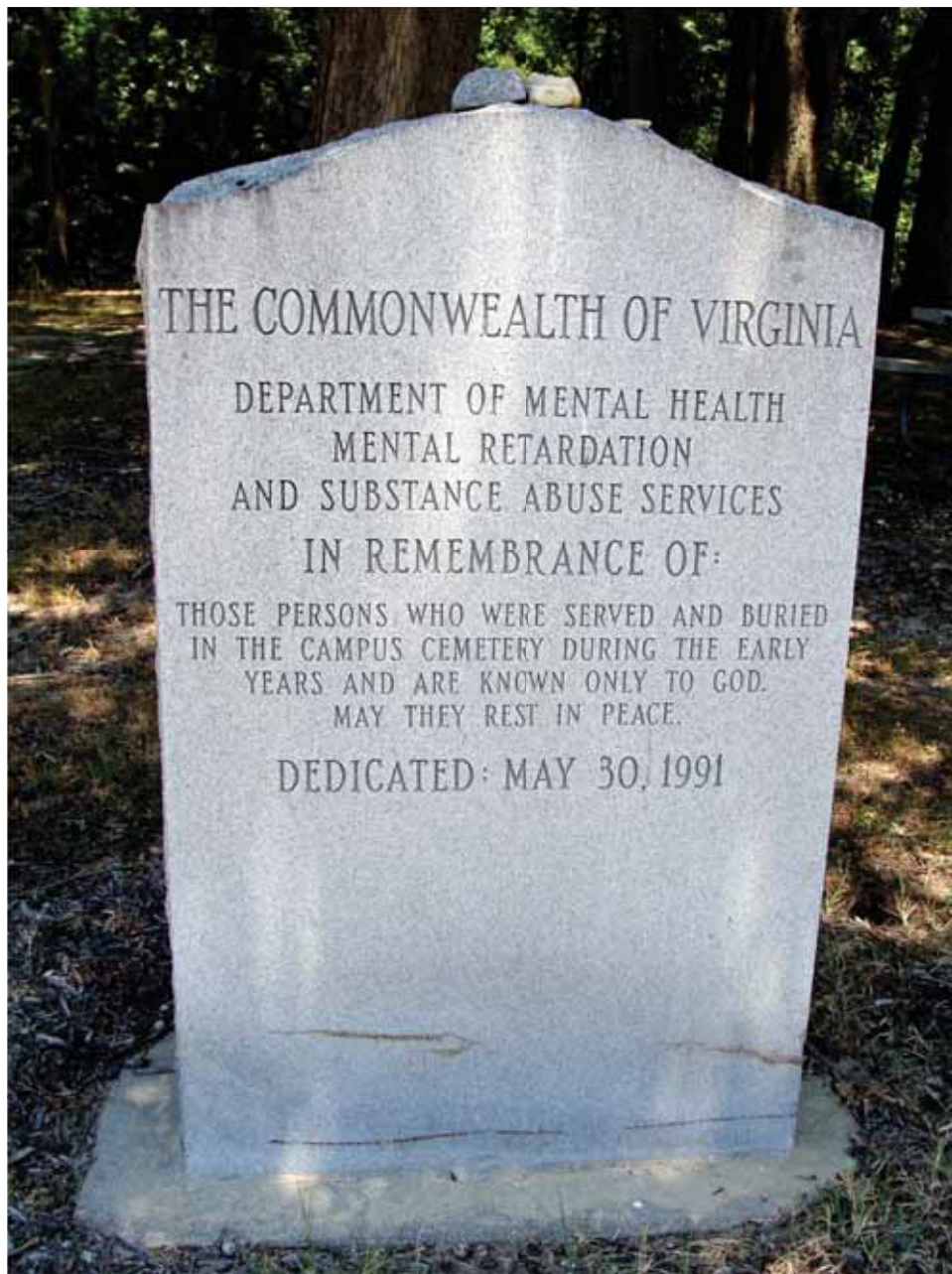


Figure 2. A starting point for commemoration at Central State Hospital. “In remembrance of those persons who were served and buried in the campus cemetery during the early years and are known only to God. May they rest in peace.” (photo by author)

L'authenticité, de l'original papier à la copie numérique

Les enjeux juridiques et archivistiques de la numérisation

Marie Demoulin et Sébastien Soyez

Résumé

La tendance actuelle semble être à la multiplication des projets de numérisation des documents papier, voire à leur élimination au profit de la seule copie numérique. Ce passage de la frontière numérique implique des mutations et soulève des questions particulières, que la présente contribution se propose d'examiner sous l'angle de l'archivistique et du droit. À travers cette approche interdisciplinaire, on s'attache à clarifier les concepts pour procéder à une qualification de la copie numérique et montrer les variations et les similitudes terminologiques autour des notions d'archive, d'original, de copie et d'authenticité. Le statut de la copie numérique est analysé, en évaluant conjointement sa valeur probante et informationnelle. Confrontant les principes à la pratique, la contribution prend la mesure du débat sur l'élimination des originaux papiers. Compte tenu des incertitudes, controverses et contradictions relevées, on souligne la nécessité d'une réforme homogène et transversale pour mieux encadrer les pratiques de numérisation et ainsi permettre une véritable reconnaissance de la copie numérique.

Auteurs

Marie Demoulin est Chercheuse, doctorante et assistante au Centre de Recherche Information, Droit et Société (CRIDS) de l'Université de Namur (FUNDP), Belgique. Membre fondatrice de FedISA Belgium. Les recherches de Marie Demoulin ont été menées dans le cadre du projet de Centre d'Expertise en Ingénierie et Qualité des Systèmes (CE-IQS), cofinancées par l'Union européenne et la Wallonie : « Le Fonds Européen de Développement Régional et la Wallonie investissent dans votre avenir ».

Sébastien Soyez est archiviste et assistant scientifique auprès de la Section « Surveillance archivistique, avis, et coordination de la collecte et de la sélection » des Archives de l'État en Belgique.

1. Introduction

Depuis plusieurs années déjà, des projets de numérisation (*digitization*) des documents papier se multiplient, tant dans le secteur privé que public, qu'ils soient apparentés à une dématérialisation totale ou partielle des processus de travail. On remarque, au cœur de ces initiatives, que la tentation est grande de forcer le passage au *tout numérique* en accélérant le processus d'élimination des documents papier. Ce phénomène n'est pas sans susciter une certaine méfiance chez les juristes et les archivistes, qui s'interrogent sur le statut et la valeur probante que l'on peut conférer à la copie numérique et s'inquiètent du sort final réservé à l'original papier. C'est pourquoi la présente contribution se propose d'examiner les enjeux de la numérisation sous l'angle du droit et de l'archivistique.

On vise ici uniquement la copie numérique, c'est-à-dire le résultat de la translation d'un document papier original vers une forme numérique, principalement suite à un processus de scanning. Nous laisserons donc de côté l'étude des originaux électroniques (*digital born*). Pourquoi un tel choix? On constate que l'environnement papier et l'environnement électronique sont souvent étudiés séparément. Les récents travaux portent généralement sur la gestion et la préservation du patrimoine numérique original, reléguant la copie numérique au second plan. À notre avis, la numérisation mérite de retenir

l'attention, étant donné l'ampleur croissante de cette pratique, mais aussi les mutations qu'impliquent le passage de la frontière numérique¹ et les questions spécifiques qui en découlent.

Sur le plan méthodologique, nous avons opté pour une approche interdisciplinaire, en croisant le regard du juriste et celui de l'archiviste. Outre qu'elle s'avère scientifiquement enrichissante, cette démarche permet de discerner de manière plus complète les enjeux du problème, la diversité et la complémentarité des points de vue et la nécessité de dégager des solutions plus cohérentes et efficaces. Il convient de préciser que le point de vue proposé est basé sur le droit de tradition civiliste, en particulier le droit belge et français, étant donné la place centrale que l'écrit y occupe sur le terrain probatoire.

Dans un premier temps, on s'attachera à clarifier les concepts pour procéder à une qualification de la copie numérique au regard des deux disciplines. Cette analyse mettra en lumière les variations et les similitudes terminologiques autour des notions d'archive, d'original, de copie et d'authenticité. Ensuite, le statut de la copie numérique sera analysé, en évaluant conjointement sa valeur probante et informationnelle. Confrontant les principes à la pratique, nous prendrons la mesure du débat sur l'élimination des originaux papiers. Compte tenu des incertitudes, controverses et contradictions relevées, nous soulignerons enfin la nécessité d'une réforme homogène et transversale pour mieux encadrer les pratiques de numérisation et ainsi permettre une véritable reconnaissance de la copie numérique.

2. Variations terminologiques

La qualification juridique et archivistique de la copie numérique passe nécessairement par l'examen des notions d'archive, d'original, de copie et d'authenticité. En l'occurrence, on observe que si ces notions sont communes aux deux disciplines, elles n'y sont pas exactement comprises de la même manière. On ne manquera pas de souligner les convergences et les divergences à cet égard et la nécessité cruciale de clarifier le sens que l'on donne aux concepts lorsqu'on touche à une question pluridisciplinaire comme la numérisation.

2.1. La notion d'archive

En Belgique, selon la législation et la terminologie archivistiques, le terme *archive* peut être décrit comme « tout document (...), quels que soient sa date, sa forme matérielle, son stade d'élaboration ou son support (...). »² A priori donc, presque tous les documents produits et reçus par un producteur d'archives³ sont considérés comme des archives.

En outre, le *stade d'élaboration* d'une archive est entendu comme la phase dans laquelle se trouve un document au moment d'être archivé. À partir de ce concept, une archive peut être envisagée sous plusieurs formes. Un original bien sûr, mais aussi une copie, une minute,⁴ un fac-similé, un brouillon, ou

¹ Nous n'examinerons pas ici le phénomène inverse, à savoir l'impression papier de documents électroniques.

² (Nous soulignons). Arrêté royal du 18 août 2010 portant exécution des articles 5 et 6 de la loi du 24 juin 1955 relative aux archives (« arrêté sur la surveillance des archives »), *M.B.*, 23 septembre 2010.

³ Un producteur d'archives est toute personne physique ou morale, publique ou privée, qui a produit, reçu et conservé des archives dans l'exercice de son activité.

⁴ Dans son acception courante administrative, la minute est la première forme d'un document (ex. dans une correspondance, la minute d'une lettre sera généralement parafée et conservée par l'expéditeur, la forme expédiée de la lettre sera conservée par le destinataire). À ne pas confondre avec la définition en droit où la minute est le nom

une traduction sont quelques exemples de stade d'élaboration d'un écrit ou d'un document. Une copie numérique peut donc recevoir la qualification d'archive.

2.2. Les notions d'original et de copie

On ne peut examiner la notion de copie sans passer par celle d'original. La notion d'original peut être diversement comprise d'une discipline à l'autre, voire au sein de la même discipline. À cet égard, pour qualifier un document d'original ou de copie, les disciplines examinées se basent sur différents critères. Ainsi, en droit de la preuve, dans les pays de tradition civiliste comme la France et la Belgique, le critère retenu est celui de la signature du document. Un autre critère, plus secondaire, est celui de la présentation du document sur son support d'origine. En archivistique, où il est plus généralement question d'identifier l'origine d'un document, c'est classiquement le critère du support qui prédomine.

2.2.1. Le critère de la signature

En droit de la preuve belge et français, l'original est synonyme d'*écrit signé*. Cette particularité mérite un mot d'explication.

Il convient d'abord de rappeler qu'en droit civil, l'écrit signé est la reine des preuves,⁵ nettement supérieure aux témoignages et aux présomptions. La preuve par écrit signée est ainsi la seule qui soit recevable en justice pour prouver un acte au-delà d'un certain montant ou pour contredire un écrit signé produit par l'autre partie.⁶ Dans ce contexte, lorsque le terme « original » est utilisé dans les dispositions du Code civil relatives à la preuve, la doctrine considère qu'il désigne l'écrit revêtu d'une signature, qu'il soit établi par des personnes privées ou par un officier public. L'original conjugue ainsi les fonctions de l'écrit et de la signature : « l'original est d'abord un écrit, c'est ensuite un écrit qui émane directement de la personne à qui on l'oppose et c'est enfin un écrit qui (...) l'identifie. »⁷ La signature est ainsi le seul critère pour pouvoir qualifier un document d'original en termes de preuve.⁸

Partant, les autres critères ne sont pas pertinents pour déterminer le caractère original d'un document sur le terrain probatoire.⁹ Ainsi, *l'original ne doit pas nécessairement être unique*.¹⁰ Au

donné à l'original d'un document, d'un acte émanant d'une juridiction ou d'un officier public (ex. acte notarié dans le cas d'un notaire, décision de justice dans le cas d'une juridiction).

⁵ Dans la hiérarchie des preuves, l'écrit signé est toutefois inférieur à l'aveu, mais il faut bien reconnaître que ce dernier n'est guère usité dans les prétoires.

⁶ Voy. l'art. 1341 C. civ. fr et b. Des exceptions sont toutefois prévues aux art. 1347 et 1348, et la preuve peuvent être administrées par tous moyens dans un certain nombre de cas. Pour plus de détails sur ces règles de preuve civile dans le domaine de l'archivage électronique, voy. M. Demoulin, « L'archivage électronique et le droit : entre obligations et précautions », dans M. Demoulin, *L'archivage électronique et le droit*, Bruxelles, Larcier, 2012, pp. 13-35.

⁷ P. Gaudrat, « Droit de la preuve et nouvelles technologies de l'information (Rapport-cadre) », dans F. Gallouédec-Genuys (dir.), *Une société sans papier? Nouvelles technologies de l'information et droit de la preuve*, Paris, La Documentation française, 1990, p. 172.

⁸ J. Larrieu, « Les nouveaux moyens de preuve : pour ou contre l'identification des documents informatiques à des écrits sous seing privé? – Contribution à l'étude juridique des notions d'écriture et de signature », *Cah. Lamy Droit de l'informatique*, 1988, H, p. 13, n° 19.

⁹ Pour plus de détails, voy. D. Gobert et E. Montero, « L'ouverture de la preuve littérale aux écrits sous forme électronique », *J.T.*, 2001, n° 6000, p. 127.

contraire, la loi exige parfois la constitution de plusieurs originaux, en autant d'exemplaires qu'il y a de parties ayant un intérêt distinct.¹¹ De plus, à moins que la loi l'exige explicitement, *l'original ne doit pas toujours être daté*,¹² même si l'absence de date pourra donner lieu à des difficultés probatoires.¹³ Enfin, *l'original ne doit pas être confondu avec son support d'origine*, sauf dans certaines hypothèses (*infra*, point 2). D'ailleurs, *l'original n'est pas une notion tributaire du papier* : un original peut désormais être un écrit électronique revêtu d'une signature électronique, les notions classiques d'écrit et de signature ayant été adaptées en ce sens par le législateur.¹⁴

A contrario, en droit, une copie est la *reproduction littérale et non signée* d'un original, *quel que soit le procédé de reproduction* (transcription manuelle, photocopie, microfilm, télécopie, numérisation, etc.).¹⁵ Même si le processus de reproduction permet de restituer à l'identique non seulement le contenu, mais également la forme de l'original, y compris les signatures des parties, il ne s'agira que d'une copie des signatures. Dans cette logique, pour qu'une copie acquière le statut d'original, elle devrait être (re)signée par son auteur.¹⁶ Naturellement, un document non signé peut également faire l'objet d'une copie, mais sa valeur probante sera encore plus limitée (*infra*, point B).

Dès lors, il est clair que la copie numérique d'un document papier signé manuscritement est juridiquement une copie. Elle ne pourrait être qualifiée d'originale¹⁷ qu'en étant revêtue, après numérisation, de la signature électronique des parties ayant signé l'original papier, afin de garantir leur identité et leur approbation du contenu. On voit d'emblée que cette façon de procéder est fastidieuse et pourrait même s'avérer impraticable.

En revanche, en archivistique, la signature n'est pas un critère exclusif pour qualifier un document d'original. Il s'agit toutefois d'un des éléments fondamentaux pour pouvoir l'authentifier (*infra*, point c, 2). Comme on va le voir, l'absence de signature n'empêche pas nécessairement un document d'être qualifié d'original.

¹⁰ Il arrive toutefois que la loi exige qu'un document original soit unique, par exemple lorsqu'il s'agit d'un titre au porteur ou d'un titre endossable, comme le connaissement maritime, dont la détention physique symbolise un droit sur les marchandises transportées par mer.

¹¹ Voy. p. ex. l'art. 1325 C. civ. b. et fr. : « Les actes sous seing privé qui contiennent des conventions synallagmatiques ne sont valables qu'autant qu'ils ont été faits en autant d'originaux qu'il y a de parties ayant un intérêt distinct ».

¹² La date n'est pas une condition de validité des actes sous seing privé, sauf lorsque la loi l'exige spécifiquement, comme pour le chèque, la lettre de change, le contrat d'assurance ou le contrat de crédit à la consommation. Par contre, elle est spécifiquement exigée pour les actes authentiques. Pour un examen de la question à propos de l'horodatage des actes électroniques, voy. M. Demoulin, « Aspects juridiques de l'horodatage des documents électroniques », M. Demoulin, D. Gobert et E. Montero, *Commerce électronique : de la théorie à la pratique*, Cahiers du CRID, n° 23, Bruxelles, Bruylant, 2003, pp. 43-68.

¹³ D. & R. Mougenot, *La preuve*, Bruxelles, Larcier, 2002, 3e éd., p. 115, n° 46; N. Verheyden-Jeanmart, *Droit de la preuve*, Bruxelles, Larcier, 1991, p. 209, n° 439 et s.

¹⁴ Voy. principalement, en Belgique, l'art. 1322, al. 2, C. civ. b., la loi du 9 juillet 2001 fixant certaines règles relatives au cadre juridique pour les signatures électroniques et les services de certification, et les art. 16 et 17 de la loi du 11 mars 2003 sur certains aspects juridiques des services de la société de l'information, *M.B.*, 17 mars 2003 *M.B.*, 29 juillet 2001; en France, les art. 1108-1 et 1316 et s. du C. civ. fr.

¹⁵ G. Cornu, *Vocabulaire juridique*, 6e éd., Paris, PUF, 1987, V° Copie et Reproduction.

¹⁶ On parle parfois ainsi de second original ou de *duplicata*.

¹⁷ Ou de second original.

2.2.2. Le critère du support d'origine

« Sous l'empire du papier, on pouvait penser, à juste titre, qu'écrit original et écrit originaire se confondent ou, en d'autres termes, que le tracé de l'écriture *sur son premier support* est une caractéristique essentielle de l'original. »¹⁸ Dans l'univers des supports traditionnels, l'original est *de facto* préservé sur son support d'origine (papier, parchemin, papyrus, etc.), auquel il est physiquement lié. Partant, le passage du support papier d'origine à un autre ferait perdre la qualification d'original à l'écrit, qui deviendrait une simple copie. Cette position a été défendue non seulement en droit, mais aussi en archivistique, où un original est un document émanant directement de son auteur, et qui constitue *l'origine et la source des reproductions* et des copies éventuelles.¹⁹

C'est pourquoi en droit belge et français, on trouve des auteurs et des décisions allant en ce sens.²⁰ Néanmoins, la doctrine majoritaire considère qu'en droit de la preuve, il est vain d'envisager la question de l'original sous l'angle du support. Peu importe si l'écriture de l'original avait été initialement apposée par les parties ou n'en était qu'une reproduction, puisque c'est bien la signature qui confère à l'écrit la qualité d'original, et non l'écriture sur un support. Ceci résout bien des controverses juridiques sur les notions d'original et de copie dans l'univers numérique, où un original électronique peut aisément changer de support durant son cycle de vie tout en restant lié à une signature électronique. Cela étant en dehors du domaine de la preuve, on trouve parfois dans la loi le terme « original » utilisé pour désigner des pièces et documents non signés. Par exemple, la loi belge sur la comptabilité des entreprises précise que « les pièces justificatives doivent être conservées, en original ou en copie (...) ». ²¹ Dans ce cas, s'agissant de documents qui ne sont pas nécessairement revêtus d'une signature (une facture, un titre de transport, un justificatif de paiement à la banque, etc.), on peut légitimement se demander si le critère de distinction entre l'original et la copie ne serait pas le support. Ce genre de conception pourrait soulever des questions pour les originaux *électroniques* non signés, lorsque la loi requiert la production d'une pièce justificative en original, mais là n'est pas notre propos.

En archivistique, le critère du support d'origine est fondamental pour vérifier le caractère original du document. Ici encore, cette conception s'avère difficile à manier dans l'environnement purement électronique. C'est pourquoi d'autres critères, plus fonctionnels, sont à l'étude pour évaluer le statut des originaux électroniques et de leurs copies, et surtout leur authenticité.²² À nouveau, ceci dépasse le cadre de la présente étude.

Dans l'hypothèse qui nous occupe, à savoir la numérisation des documents papier, la situation est claire : il s'agit simplement d'un original papier reproduit sous forme numérique. Dès lors, en droit comme en archivistique, le fichier numérique ainsi obtenu sera toujours qualifié de copie, soit par la perte de la signature d'origine, soit en raison du changement de support.

¹⁸ D. Gobert et E. Montero, « L'ouverture de la preuve littérale aux écrits sous forme électronique », *J.T.*, 2001, n° 6000, p. 127.

¹⁹ Définitions reprises du glossaire présenté sur le Portail international archivistique francophone (PIAF), <http://www.piaf-archives.org>, consulté le 28/08/2012.

²⁰ Liège, 20 juin 1978, *Jur. Liège*, 1978-1979, p. 17. Voy. aussi Douai, 25 octobre 1966 et T.G.I. Bayonne, 5 juillet 1976, cités par M. Van Quickenborne, « Quelques réflexions sur la signature des actes sous seing privé », note sous Cass., 28 juin 1982, *R.C.J.B.*, 1985, p. 91, n° 32, note 100. Voy. aussi X. Malengreau, « Le droit de la preuve et la modernisation des techniques de rédaction, de reproduction et de conservation des documents », *Ann. dr. Louvain*, 1981/2, p. 115.

²¹ Art. 6 de la loi belge du 17 juillet 1975 relative à la comptabilité des entreprises.

²² Voy. not. les importants travaux du projet InterPARES (www.interpares.org).

Il est singulier de constater qu'en archivistique, la notion de copie est déclinée sous diverses variantes. On y retrouve des concepts voisins comme celui de copie d'utilisation (ou de consultation), de copie de sécurité ou encore de copie de substitution.²³ Ces termes existaient déjà avant l'avènement du numérique et étaient appliqués notamment pour le microfilmage. La *copie d'utilisation* est une copie effectuée généralement par le service d'archives lui-même dans le but de faciliter la communication et d'éviter que l'original papier soit malmené par des consultations répétées (ex. copies des registres paroissiaux). La *copie de sécurité* est effectuée dans le but de conserver une copie d'un document au cas où l'original serait détérioré ou détruit accidentellement.²⁴ La *copie de substitution* est une copie d'un document original, en vue de le remplacer et de pouvoir le détruire. À cet égard, ces qualifications peuvent sans difficulté s'appliquer à la copie numérique d'un document papier.

2.3 La notion d'authenticité

L'authenticité est une notion fréquemment utilisée en droit et en archivistique, et plus précisément en diplomatique. Il convient de distinguer l'authenticité conférée *a priori* à des écrits spécifiques et l'authenticité vérifiée *a posteriori* sur tout type de document.

2.3.1. L'authenticité conférée a priori

Dans les pays de tradition civiliste, l'authenticité en droit est une qualité spéciale conférée *a priori* à certains types d'actes juridiques, appelés actes authentiques. « L'authenticité est le caractère de vérité et de force qui s'attache aux actes de l'autorité publique. »²⁵ « Dans chaque État, cette qualité est conférée par le Pouvoir qui l'institutionnalise et l'organise à son gré en fonction de son système juridique. »²⁶ À titre d'exemple, sont des actes authentiques les actes notariés, les exploits d'huissier, les jugements ou encore les actes de l'état civil.²⁷

Rappelons que parmi les actes juridiques, on distingue les actes sous seing privé et les actes authentiques. Au sens matériel, les actes juridiques sont des titres, des instruments, bref des écrits, rédigés en vue de faire preuve.²⁸ L'acte sous seing privé est un acte établi entre particuliers et revêtu de leur signature.²⁹ L'acte authentique est un acte officiel, reçu ou dressé par un officier public³⁰ dûment habilité,

²³ R. Petit., D. Van Overstraeten, H. Coppens et J. Nazet, *Terminologie archivistique en usage aux Archives de l'État en Belgique, vol. I, Gestion des archives*, Série Miscellanea Archivistica. Manuale (16), Bruxelles, Archives générales du Royaume et Archives de l'État dans les provinces, 1994.

²⁴ On parle aussi de copie miroir, notion ancienne utilisée actuellement comme pierre angulaire des politiques de sécurité informatique.

²⁵ J. Demblon, P. Harmel, M. Renard-Declairfayt, J.-F. Taymans, *L'acte notarié*, Rép. Not. T. XI, L. VII, Bruxelles, Larcier, 2002, p. 103, n° 13.

²⁶ D. & R. Mougenot, *La preuve*, Bruxelles, Larcier, 2002, 3e éd., n° 85, p. 149.

²⁷ Voy. les exemples cités par D. & R. Mougenot, *La preuve*, Bruxelles, Larcier, 2002, 3e éd., n° 86, p. 149. Voy. aussi la typologie dressée par I. de Lamberterie, « Réflexions sur l'établissement et la conservation des actes authentiques », Les actes authentiques électroniques — Réflexion juridique prospective, Paris, La documentation française, 2002, p. 33, également disponible à l'adresse [http://www.interpares.org/display_file.cfm?doc=ip2\(delamberterie\)_reflexions_etablissement_et_conservation_actes_authentiques.pdf](http://www.interpares.org/display_file.cfm?doc=ip2(delamberterie)_reflexions_etablissement_et_conservation_actes_authentiques.pdf), p. 8.

²⁸ N. Verheyden-Jeanmart, *Droit de la preuve*, Bruxelles, Larcier, 1991, p. 194, n° 396.

²⁹ Normalement, l'acte sous seing privé ne doit répondre à aucune exigence de forme particulière. Cependant, il est de plus en plus fréquent que la rédaction de certains types de contrat (contrat de crédit, contrat de travail, contrat de

dans les formes requises par la loi.³¹ Ces formes varient d'un acte authentique à l'autre et sont précisées dans les textes de loi propres à chacun d'eux. Parmi les formes récurrentes, on peut citer, entre autres, la signature de l'officier public, son sceau, la date de l'acte et l'apposition de mentions ou de formules spécifiques. Trois effets sont principalement attachés à l'acte authentique :³² il jouit d'une force probante supérieure jusqu'à inscription de faux,³³ sa date est certaine³⁴ et il a force exécutoire.³⁵ Et encore, seules certaines mentions de l'acte sont couvertes par l'authenticité, à savoir l'origine de l'écriture des parties et de l'officier public instrumentant, ainsi que toutes les constatations faites personnellement par l'officier public au moment d'instrumenter et dans les limites de sa compétence et de sa mission.³⁶

Il convient encore de noter que l'authenticité ne semble pas être une qualité propre à l'original. Ainsi, dans le cas des actes notariés, il est question de *copies authentiques*³⁷ (les grosses et les expéditions³⁸) délivrées par le notaire et certifiées conformes à l'original (la minute, conservée par le notaire et dont il ne peut se dessaisir).

À cet égard, on note qu'une copie numérique d'un acte notarié papier peut être qualifiée de « copie authentique » en droit français, lorsqu'elle est réalisée dans certaines conditions réglementaires.³⁹ Le décret relatif aux actes établis par les notaires prévoit en effet que « Les copies authentiques sont établies soit sur support papier, soit sur support électronique, quel que soit le support initial de l'acte. »⁴⁰ En droit belge, il est prévu que « Tous les actes notariés reçus sous forme dématérialisée, ainsi qu'une copie dématérialisée de tous les actes qui sont reçus sur support papier, sont conservés dans une Banque des actes notariés. (...) La Banque des actes notariés a la valeur de *source authentique* pour les actes qui y

consommation...) soit assortie d'exigences légales de forme, comme la datation, certaines mentions obligatoires, etc. Ces exigences de forme n'érigent toutefois pas le contrat au rang d'acte authentique.

³⁰ Notamment, un notaire, un juge, un greffier, un huissier de justice, un officier de l'État civil, un officier de police judiciaire... Voy. N. Verheyden-Jeanmart, *Droit de la preuve*, Bruxelles, Larcier, 1991, p. 211, n° 445.

³¹ Voy. l'art. 1317, al. 1, C. civ. b. et fr. : « L'acte authentique est celui qui a été reçu par officiers publics ayant le droit d'instrumenter dans le lieu où l'acte a été rédigé, et avec les solennités requises ».

³² I. de Lamberterie, « Réflexions sur l'établissement et la conservation des actes authentiques », Les actes authentiques électroniques — Réflexion juridique prospective, Paris, La documentation française, 2002, p. 28, également disponible à l'adresse

[http://www.interpares.org/display_file.cfm?doc=ip2\(delamberterie\)_reflexions_etablissement_et_conservation_actes_authentiques.pdf](http://www.interpares.org/display_file.cfm?doc=ip2(delamberterie)_reflexions_etablissement_et_conservation_actes_authentiques.pdf), p. 4.

³³ Autrement dit, les mentions de l'acte qui sont couvertes par l'authenticité ne peuvent être attaquées qu'en recourant à une procédure particulière d'inscription de faux, compliquée et coûteuse, visant à démontrer que l'acte produit est en réalité un faux. Voy. N. Verheyden-Jeanmart, *Droit de la preuve*, Bruxelles, Larcier, 1991, p. 222, n° 469 et s.

³⁴ C'est-à-dire opposable aux tiers.

³⁵ « La force exécutoire d'un acte est la possibilité de recourir à l'exécution forcée de cet acte, lorsqu'il a été revêtu de la formule exécutoire ». Voy. N. Verheyden-Jeanmart, *Droit de la preuve*, Bruxelles, Larcier, 1991, p. 196, n° 403.

³⁶ D. & R. Mougenot, *La preuve*, Bruxelles, Larcier, 2002, 3e éd., p. 155, n° 93.

³⁷ On retrouve cette expression dans le décret français n° 71-941 du 26 novembre 1971 relatif aux actes établis par les notaires, art. 33.

³⁸ « L'expédition est la copie entière, littérale, certifiée conforme par le notaire, mais dépourvue de la formule exécutoire, de l'acte restant en mains du notaire ». « La grosse est une expédition munie de la formule exécutoire » Voy. J. Demblon, P. Harmel, M. Renard-Declairfayt, J.-F. Taymans, *L'acte notarié*, Rép. Not. T. XI, L. VII, Bruxelles, Larcier, 2002, pp. 432 et 434.

³⁹ Voy. l'art. 37 du décret français n° 71-941 du 26 novembre 1971 relatif aux actes établis par les notaires, modifié par le décret n° 2005-973 du 10 août 2005.

⁴⁰ Voy. l'art. 33 du décret français n° 71-941 du 26 novembre 1971 relatif aux actes établis par les notaires, modifié par le décret n° 2005-973 du 10 août 2005.

sont enregistrés.»⁴¹ Cela étant dit, il n'est pas certain que le terme *authentique* ait ici la même signification. Il semble utilisé pour désigner, d'un côté, le caractère authentique d'un acte (reçu par l'officier public compétent selon les formalités requises), de l'autre, le caractère authentique de son origine. Quoi qu'il en soit, ces copies authentiques n'ont pas tout à fait la même force probante que l'acte authentique original, comme on le verra (*infra*, point B).

En résumé, en droit, l'authenticité se confère *a priori*,⁴² au moment de l'élaboration de certains types d'actes, moyennant le respect de certaines formes spécifiques et l'intervention d'un officier public compétent. L'authenticité est ainsi le propre des actes authentiques, à l'exclusion des autres écrits. Cette qualité leur procure une valeur juridique particulière, notamment comme moyen de preuve. Pour autant, une autre signification, plus large et plus courante, est également attribuée à la notion d'authenticité en droit. Est authentique l'objet ou le document, quel qu'il soit, dont l'auteur ou l'origine ont pu être vérifiés *a posteriori*.⁴³ Cette seconde conception juridique de l'authenticité rejoint celle qui domine la discipline archivistique, et plus précisément la diplomatique.

2.3.2. L'authenticité vérifiée *a posteriori*

La recherche de l'authenticité renvoie, en archivistique, à la notion de diplomatique. La diplomatique est une science auxiliaire à l'archivistique et à l'histoire. Elle étudie les actes écrits en eux-mêmes et, par extension, tous les documents d'archives, d'après leur forme, leur genèse et leur tradition, et en établit la typologie.⁴⁴ Pour simplifier, on peut dire que les techniques utilisées en diplomatique visent à s'interroger sur la crédibilité de l'écrit que l'on analyse.

L'authenticité diplomatique renvoie à la conclusion selon laquelle un document est bien ce qu'il prétend être, après qu'il ait subi avec succès une analyse et une critique de sa forme, examen qui aura mis en évidence que le discours émane bien de la personne qui apparaît comme l'auteur, et que ce discours a bien été établi et validé à la date affichée ou suggérée dans le document. Ainsi, un document authentique est un document dont l'exactitude, la véracité ne peuvent être contestées.

Cette authenticité s'apprécie à l'égard de n'importe quel type de document, quel que soit son auteur ou la forme de sa rédaction, qu'il s'agisse d'un original ou d'une copie, d'un écrit signé ou non signé, d'un acte « authentique » (au sens juridique premier) ou d'un acte sous seing privé.

Pour que la diplomatique puisse s'exercer, il faut nécessairement que le document à authentifier soit fini et daté, faute de quoi l'exigence d'authenticité n'a pas lieu d'être. Il est en effet impossible de vérifier l'auteur ou la date d'un document qui n'est pas fixé dans le temps, ou de vérifier l'auteur d'un contenu qui n'a pas d'auteur.⁴⁵

La vérification de l'authenticité d'un document est facilitée quand il s'agit de sa rédaction définitive et que l'on peut identifier son auteur par le biais d'une signature manuscrite. D'autres éléments vont concourir à conforter son authenticité : des inscriptions complémentaires, appelées en diplomatique les

⁴¹ Voy. les art. 18 et 20 de la loi belge du 16 mars 1803 contenant organisation du notariat modifiée par la loi du 6 mai 2009 (*M.B.*, 19 mai 2009). Ces dispositions entreront, en vigueur à une date à fixer par le Roi.

⁴² G. Cornu, *Vocabulaire juridique*, Paris, PUF, 1996, 6e éd., V° Authenticité (2), Authentification (1) et Authentique (2).

⁴³ G. Cornu, *Vocabulaire juridique*, Paris, PUF, 1996, 6e éd., V° Authenticité (1), Authentification (2) et Authentique (1).

⁴⁴ Glossaire du PIAF, *op.cit.*

⁴⁵ M.-A. Chabin, *Histoire de dates. Réflexion de diplomatique numérique*, Montréal, École de Bibliothéconomie et des Sciences de l'Information (EBSI), 1er novembre 2011.

« mentions marginales », comme le numéro et la date d'enregistrement du document, qui sont apportées lors de sa diffusion. En outre, une analyse graphologique des écritures ou des expertises chimiques sur le support, les timbres ou autres sceaux éventuels dont il est revêtu concourent encore à en établir le caractère authentique.

Lorsqu'il est établi que l'on est face à un faux, il est intéressant de constater que le document ne perd pas systématiquement toute valeur archivistique justifiant sa conservation, à la différence du droit qui va l'écarter catégoriquement des débats. Le faux peut en effet présenter une valeur patrimoniale propre, non plus en tant que reproduction fidèle de l'original, mais au contraire en tant que pièce falsifiée, notamment si l'on peut identifier le faussaire et les enjeux historiques d'une telle manipulation. On pourrait presque parler à cet égard d'un « authentique faux ».

Par ailleurs, on relève qu'une copie peut être authentique lorsqu'elle est une copie intégrale d'un document délivré notamment par un service d'archives public, soit sous forme de photographie, de photocopie ou de copie numérique, soit sous forme de transcription manuscrite, et accompagnée des signes légaux d'authentification.⁴⁶ Comme en droit, l'authenticité diplomatique n'est pas forcément liée au statut original d'un document, mais bien, dans le cas d'une copie, à la conformité de la copie avec l'original. Il est par contre difficile d'authentifier de manière absolue une copie quand on ne dispose plus de l'original.

Dans cette optique, on constate que la diplomatie procède de la même logique que les vérifications d'écriture et de signature, et plus largement les expertises auxquelles il peut être procédé pour vérifier la véracité d'un document produit comme moyen de preuve devant le juge. Cette convergence n'a rien d'étonnant dans la mesure où, historiquement, la diplomatie était utilisée en justice pour authentifier les actes anciens, afin de déterminer l'existence de droits de propriété ou de privilèges.⁴⁷ À cet égard, certains ont déjà plaidé, avec justesse, pour un rapprochement de la diplomatie numérique (*Digital Diplomatics*) et des méthodes d'expertise légale des documents électroniques (*Digital Records Forensics*).⁴⁸ L'expertise des documents numérisés devrait en faire partie, en tenant compte des particularités de leur processus de création. En outre, à l'ère numérique, il est nécessaire que la diplomatie se modernise et développe de nouveaux critères, plus fonctionnels, non seulement pour authentifier *a posteriori* l'origine et le contenu des documents électroniques (natifs ou numérisés), mais aussi pour intervenir *a priori*, de manière prospective, afin que les documents destinés à être conservés soient munis de toutes les garanties d'authenticité dès leur conception.⁴⁹

3. Valeur de la copie numérique d'un document papier

Comme on l'a vu, le fichier électronique obtenu après avoir scanné un document papier n'est qu'une copie numérique de celui-ci. Il convient dès lors de s'interroger sur le statut, la valeur de cette copie numérique, au regard du droit et de l'archivistique.

⁴⁶ La notion de copie authentique est présentée à l'article 3 de la loi belge du 24 juin 1955 relative aux archives (*M.B.*, 12 août 1955), où l'on précise que « *Les expéditions* (ndr : en d'autres mots les copies littérales d'un acte, délivrées en bonne forme par l'officier public, dépositaire de l'original, c'est-à-dire les Archives de l'État) *ou extraits sont délivrés par les conservateurs des archives, signés par eux et munis du sceau du dépôt; ils font ainsi foi en justice* ».

⁴⁷ Voy. L. Duranti, « Diplomats », *Encyclopedia of Library and Information Sciences*, 3e éd., 2010, 1 : 1, p. 1593.

⁴⁸ Voy. L. Duranti, « From Digital Diplomats to Digital Records Forensics », *Archivaria*, n° 68, 2009, p. 39 et s.

⁴⁹ À propos de cette diplomatie moderne, par contraste avec la diplomatie classique, voy. L. Duranti, « Diplomats », *Encyclopedia of Library and Information Sciences*, 3e éd., 2010, 1 : 1, p. 1594 et s.

3.1. La force probante de la copie numérique

On s'attache essentiellement ici à évaluer la valeur juridique de la copie numérique au regard du droit de la preuve, d'une perspective civiliste.⁵⁰ À cet égard, le principe qui prévaut est que la copie (numérique ou non) n'a pas la force probante d'un original. Ce n'est que dans des hypothèses limitées qu'une force probante presque équivalente à l'original est accordée à certaines copies, sans franchir le pas d'une équivalence totale.

3.1.1. Le principe : force probante inférieure à l'original

En Belgique et en France, les articles 1334 et suivants du Code civil définissent le régime probatoire des copies. Or, ces dispositions ne visent que les copies d'actes authentiques qui, seules et à certaines conditions, se voient reconnaître une force probante dans la loi, parce qu'elles émanent de l'autorité publique. Les copies réalisées par les particuliers ne sont en principe pas visées par le Code civil, même si cette question demeure controversée.⁵¹

Quoi qu'il en soit, le principe reste le même pour toutes les copies, que l'on applique l'article 1334 du Code civil ou non :⁵² on peut toujours contester une copie et réclamer la production de l'original pour lever un doute sur la véracité de son contenu.⁵³

Le vrai problème se pose lorsqu'on n'a pas conservé l'original : dans ce cas, quel crédit accorder à la copie? Lorsqu'il s'agit de copies d'actes authentiques, leur force probante est déterminée par les articles 1335 et s. du Code civil, dont nous n'examinons pas le détail ici (*infra*, point 2). Quant aux copies d'actes sous seing privé (ou d'actes non signés, d'ailleurs), face au silence du législateur, il appartient au juge de se prononcer sur leur valeur probante. À notre connaissance, il n'existe pas de jurisprudence belge ou française sur la valeur probante des documents scannés. Toutefois, il nous semble que l'on peut raisonner à partir de la jurisprudence développée en matière de photocopie.⁵⁴

Si l'original n'a pas été conservé, la copie n'aura pas beaucoup de valeur probante, mais cela ne signifie pas qu'elle n'aura pas de valeur du tout. On pourrait ainsi considérer la copie numérique d'un document papier comme un commencement de preuve par écrit.⁵⁵ Un commencement de preuve par écrit est, selon l'article 1347 du Code civil, un écrit qui n'est pas signé, mais qui émane bien de celui à qui on l'oppose et qui rend vraisemblable le fait allégué en justice. Or, la copie numérique peut être considérée comme un écrit électronique. S'il s'agit d'une copie d'un document signé, ou comportant des mentions manuscrites, ou d'autres marques indiquant son origine, l'on pourrait considérer qu'il émane bien de leur auteur. Ce raisonnement ne vaut, bien entendu, que si la fidélité de la copie par rapport à l'original ne peut raisonnablement être mise en doute. On ne saurait trop insister sur l'importance capitale d'éléments

⁵⁰ Pour plus de détails, voy. M. Demoulin, « L'archivage électronique et le droit : entre obligations et précautions », dans M. Demoulin, *L'archivage électronique et le droit*, Bruxelles, Larcier, 2012, pp. 13-35.

⁵¹ Certains auteurs estiment ainsi que l'article 1334 C. civ. pourrait s'appliquer aux actes sous seing privé. Sur cette controverse, voy. D. & R. Mougenot, *La preuve*, Bruxelles, Larcier, 2002, 3e éd., pp. 248-249, n° 188, et les réf. citées.

⁵² Ibidem.

⁵³ Art. 1334 C. civ. b. et fr. : « Les copies, lorsque le titre original subsiste, ne font foi que de ce qui est contenu au titre, dont la représentation peut toujours être exigée ».

⁵⁴ Pour une étude en la matière, voy. D. Mougenot, « Le régime probatoire de la photocopie et du téléfax », in *La preuve*, Liège, Formation permanente CUP, Liège, 2002, p. 229-268.

⁵⁵ Voy., par analogie, le raisonnement développé par D. Mougenot, « Le régime probatoire de la photocopie et du téléfax », dans *La preuve*, Liège, Formation permanente CUP, Liège, 2002, p. 245-252, n° 14 et s., et les réf. citées.

comme les métadonnées du document, la procédure de scanning et d'archivage électronique, la personne sous le contrôle de laquelle la numérisation a eu lieu, etc., pour convaincre le juge que le processus de numérisation n'a pas été entaché de fraude. En outre, le commencement de preuve par écrit n'est pas une preuve qui se suffit à elle-même. Pour emporter la conviction du juge, il faudra encore le compléter par des témoignages ou des présomptions. Toutefois, l'avantage indéniable est la recevabilité de ce mode de preuve, même dans les cas où, normalement, le juge ne pourrait accepter qu'une preuve par écrit signée.⁵⁶

Lorsque la copie numérique ne réunit pas les conditions d'un commencement de preuve par écrit, elle sera tout au plus considérée comme une simple présomption. Cependant, à la différence du commencement de preuve par écrit, une présomption n'est pas recevable en justice lorsque la loi exige une preuve par écrit signée. Le juge n'acceptera de la prendre en considération que dans les cas où la preuve peut être administrée par tous moyens,⁵⁷ mais il appréciera souverainement son caractère convainquant.

On ajoute encore qu'en principe, la copie certifiée conforme ne jouit pas d'une force probante supérieure, sauf lorsque la loi l'indique spécifiquement. Cependant, ce type de copie s'avérera plus convainquant pour le juge, mais seulement au titre de commencement de preuve par écrit ou de présomption.⁵⁸

On voit que le statut de commencement de preuve par écrit ou de présomption est largement inférieur à celui d'écrit signé, de sorte que la copie numérique ne jouit pas de la plus grande force probante en l'état actuel des textes. Étant donné les importantes incertitudes à cet égard, il convient donc de bien mesurer les risques juridiques avant d'envisager la destruction de l'original papier après sa numérisation.

3.1.2. Exceptions : force probante (quasi) équivalente à l'original

Cet état des lieux doit cependant être nuancé, dans la mesure où il existe des cas où la copie aura la même force probante que l'original... ou presque.

3.1.2.1. Les copies non contestées

D'abord, rappelons que si sa conformité à l'original n'est pas contestée par l'autre partie, la copie fera pleine preuve de son contenu.⁵⁹ On relève ainsi que certaines décisions ont déjà accordées à la photocopie une force probante particulière :⁶⁰ si le défendeur n'en conteste pas le contenu ou en avoue la sincérité par son attitude, elle peut se voir attribuer « le même caractère de véracité que les lettres originales elles-mêmes »⁶¹ ou, du moins, dispenser de la production de l'original.⁶² Cela ne signifie pas que la copie a la

⁵⁶ Le recours au commencement de preuve par écrit est en effet recevable par exception à la règle qui exige une preuve par écrit signée (art. 1341 et art. 1347 C. civ. b. et fr.).

⁵⁷ On songe par exemple aux litiges en matière commerciale, ou au cas où l'original aurait été perdu ou détruit par cas fortuit ou de force majeure (art. 1348, 4°, C. civ.).

⁵⁸ D. & R. Mougenot, *La preuve*, Bruxelles, Larcier, 2002, 3e éd., pp. 249-250, n° 189.

⁵⁹ Voy. D. Mougenot, « Le régime probatoire de la photocopie et du télécopie », in *La preuve*, Liège, Formation permanente CUP, Liège, 2002, p. 254, n° 26, et les réf. citées.

⁶⁰ Ph. Malaurie, note sous Cass. fr. (civ.), 21 avril 1959, *D.*, 1959, p. 522.

⁶¹ Cass. fr. (civ.), 20 juil. 1953, *J.C.P.*, G, 1953, note J. Savatier, qui considère cette formule comme équivoque et dangereuse.

même force probante que l'acte original, puisqu'elle pourra toujours être contestée par toutes voies de droit, alors que l'acte original ne peut être contesté que par la production d'un autre écrit signé (pour les actes sous seing privé⁶³) ou par une procédure d'inscription de faux (pour les actes authentiques).

Or, on sait qu'en pratique, devant les tribunaux, la plupart des parties présentent à la cause des photocopies de documents, généralement sans que cela pose le moindre problème. Sauf en cas de doute sérieux sur la conformité de la copie, il est rare qu'une partie exige la production de l'original. Il devrait en être de même pour la plupart des documents scannés, mais toujours sous réserve du droit pour l'autre partie de réclamer l'original, qui plane comme une épée de Damoclès au-dessus de la copie.

3.1.2.2. Certaines copies d'actes authentiques

On note également, sans entrer dans les détails, que certaines copies authentiques, comme les grosses, les expéditions, et les documents qui y sont assimilés, se voient reconnaître la force probante de l'acte authentique. D'autres copies d'actes authentiques ne valent cependant que commencement de preuve par écrit.⁶⁴

En Belgique, en ce qui concerne les copies numériques d'actes notariés reçus sur support papier, elles devront être conservées dans une Banque des actes notariés gérée par la Chambre nationale des notaires et elles auront la même valeur probante que la première expédition de la minute sur support papier⁶⁵

On note cependant que la force probante de ces copies authentiques n'est pas tout à fait celle des actes authentiques originaux, puisque pour les contester, il ne faut pas tenter une procédure en inscription de faux : il suffit de demander la production de l'original (art. 1334 C. civ.). Ce n'est que si l'original n'existe plus que ces copies authentiques sont totalement assimilées à l'acte authentique original (art. 1335 C. civ.).

3.1.2.3. La copie fidèle et durable en France

En 1980, le législateur français a modifié le Code civil pour y ajouter une disposition prévoyant que lorsqu'une partie ou le dépositaire n'a pas conservé le titre original, mais dispose d'une copie qui en est la reproduction fidèle et durable, cette copie peut être présentée en justice à la place de l'original.⁶⁶ La majorité des auteurs estime que le législateur dispense ainsi les parties de conserver le titre original, qui peut être volontairement détruit après copie, sans perdre tout moyen de preuve.⁶⁷ Mais la copie fidèle et

⁶² Cass. fr. (civ.), 21 avril 1959, *D.*, 1959, p. 521, note Ph. Malaurie; Cass., 21 février 1964, *Pas.*, 1964, I, p. 664; Cass. fr. (civ.), 30 avril 1969, *J.C.P.*, G, 1969, note M.A.

⁶³ Voy., pour rappel, l'art. 1341 C. civ. b. et fr.

⁶⁴ Voy. les art. 1335 et s. C. civ. b. et fr.

⁶⁵ Voy. l'art. 18 de la loi belge du 16 mars 1803 contenant organisation du notariat modifiée par la loi du 6 mai 2009 (*M.B.*, 19 mai 2009). Ces dispositions entreront, en vigueur à une date à fixer par le Roi.

⁶⁶ Art. 1348, al. 2, C. civ. fr.

⁶⁷ Ph. Jestaz, Commentaire de la loi n° 80-525 du 12 juillet 1980, *Rev. trim. dr. civ.*, 1980, p. 821; J. Viatte, « La preuve des actes juridiques – Commentaire de la loi n° 80-525 du 12 juillet 1980, *Gaz. pal.*, 1980, p. 582; F. Chamoux, « La loi du 12 juillet 1980 : une ouverture sur de nouveaux moyens de preuve », *JCP*, G, 1981, I 3008, n° 23 et s.; I. Pottier, note sous Cass. fr. (civ.), 30 juin 1993, *Gaz. pal.* 1993, p. 469; J. Ghestin, G. Goubeaux et M. Fabre-Magnan, *Traité de droit civil – Introduction générale*, 4e éd., Paris, L.G.D.J., 1994, p. 651, n° 670; J. Huet et H. Maisl, *Droit de l'informatique et des télécommunications*, Paris, Litec, 1989, p. 665, n° 594.

durable ne se voit pas accorder la même valeur que l'écrit original : elle est recevable en justice, mais peut être contestée par tous les moyens.⁶⁸

En outre, seule la copie « fidèle et durable » bénéficie d'une telle reconnaissance. Or, la notion de copie durable est définie dans le Code civil comme « toute reproduction indélébile de l'original, qui entraîne une modification irréversible de son support. »⁶⁹ Cette définition semble se cantonner à des supports matériels, comme le microfilm, voire le CD-ROM, au détriment des procédés logiciels ou organisationnels permettant de garantir l'intégrité du contenu et au mépris des réalités de la pratique, étant donné que le maintien de la pérennité et de la lisibilité du document pourrait nécessiter des migrations régulières de support. Dès lors, à moins que sa fidélité soit certaine et qu'elle soit enregistrée de manière « indélébile » entraînant une « modification irréversible de son support », la copie numérique ne peut être considérée comme une copie fidèle et durable aux yeux du droit français.

3.1.2.4. Certaines copies, dans des secteurs spécifiques

On relève encore qu'à titre exceptionnel, on trouve dans la législation des dispositions éparpillées octroyant force probante à certaines copies, y compris les copies numériques.

Certaines de ces dispositions dérogatoires concernent les copies réalisées par des organismes privés, mais la plupart visent les copies faites sous le contrôle d'organismes publics.⁷⁰ Ainsi, les copies de documents réalisées par les entreprises d'assurance et les établissements de crédit « font foi comme les originaux sauf preuve contraire. »⁷¹ De même, les copies faites par les institutions de sécurité sociale ont « valeur probante jusqu'à preuve du contraire » à condition que la procédure d'enregistrement, de conservation et de reproduction des documents soit dûment agréée par le ministre compétent.⁷² Il peut s'agir de copies de tous types de documents, et non uniquement de copies d'écrits signés. En leur accordant « valeur probante jusqu'à preuve du contraire », le législateur renverse la charge de la preuve. En effet, celui qui conteste la fiabilité de la copie ne peut plus se contenter de demander la production de l'original : il lui appartient d'en établir le caractère non fiable, avec toute la difficulté d'une telle preuve négative. Cette preuve contraire pourra toutefois être administrée par tous les moyens, et non uniquement

⁶⁸ Ph. Jestaz, Commentaire de la loi n° 80-525 du 12 juillet 1980, *Rev. trim. dr. civ.*, 1980, p. 821; F. Chamoux, « La loi du 12 juillet 1980 : une ouverture sur de nouveaux moyens de preuve », *JCP*, G, 1981, I 3008, n° 15; J. Huet et H. Maisl, *Droit de l'informatique et des télécommunications*, Paris, Litec, 1989, p. 666, n° 594.

⁶⁹ Art. 1348, al. 2, C. civ. fr.

⁷⁰ Voy. not., en Belgique : Arrêté royal du 15 mars 1999 relatif à la valeur probante, en matière de sécurité sociale et de droit du travail, des informations échangées, communiquées, enregistrées, conservées ou reproduits par les services ministériels et les parastataux du Ministère de l'Emploi et du Travail, *M.B.*, 5 juillet 1999; Arrêté royal du 9 janvier 2000 relatif à la force probante des informations utilisées par l'Administration des Pensions pour l'application de la législation dont elle est chargée, *M.B.*, 24 février 2000; Arrêté royal du 26 avril 2007 modifiant l'arrêté royal du 27 avril 1999 relatif à la force probante des données enregistrées, traitées, reproduites ou communiquées par les dispensateurs de soins et les organismes assureurs, *M.B.*, 18 juin 2007; Art. 13 de la Loi du 21 août 2008 relative à l'institution et à l'organisation de la plate-forme eHealth, *M.B.*, 13 octobre 2008.

⁷¹ Art. 196 de la loi belge du 17 juin 1991 portant organisation du secteur public du crédit et de la détention des participations du secteur public dans certaines sociétés financières de droit privé.

⁷² Arrêté royal du 22 mars 1993 portant organisation du secteur public du crédit et de la détention des participations du secteur public dans certaines sociétés financières de droit privé, pris en application de l'art. 18 de la loi du 4 avril 1991.

par la production d'un écrit signé. Dès lors, techniquement, la copie d'un acte signé n'aura pas la même force probante que l'original, puisqu'elle peut être contredite par davantage de moyens de preuve.⁷³

On relève toutefois l'existence de quelques rares textes, plus anciens d'ailleurs, reconnaissant que les copies effectuées par certains organismes sont purement et simplement équivalentes à l'original.⁷⁴ Cela signifie que la copie d'un acte signé ne pourra être contredite par témoignages ou présomptions, mais uniquement par la production d'un autre écrit signé, ce qui lui donne véritablement une force probante égale à l'original signé. Cela étant, il est intéressant de constater que plusieurs textes reconnaissant une telle équivalence pure et simple⁷⁵ ont par la suite été modifiés pour y ajouter la possibilité d'une preuve contraire, ce qui dénote peut-être un recul du législateur à l'égard de la copie.

Il existe également des autorisations sectorielles de *conserver* un document sous forme de copie, mais elles sont très limitatives. Par exemple, en Belgique, les factures,⁷⁶ les pièces justificatives de comptabilité⁷⁷ ou les documents sociaux⁷⁸ peuvent être conservés sous forme de copie ou d'original. Ces règles n'octroient toutefois pas explicitement une force probante particulière à ces copies.

Il est loisible au législateur de déroger aux règles du Code civil relatives à la force probante de la copie,⁷⁹ voire de donner délégation au pouvoir réglementaire pour le faire. On note cependant qu'en Belgique, ces textes peuvent poser problème dans le secteur public, au regard de la loi sur les archives, qui interdit de procéder à une destruction d'archives sans l'autorisation préalable de l'Archiviste général du Royaume,⁸⁰ comme on va le voir.

3.2. La valeur archivistique

Dans la sphère archivistique, le statut de copie (numérique ou papier) ne dénature pas en soi la valeur de l'information qui y est enregistrée, pour autant qu'on ait pu authentifier son contenu (*supra*, point A, c, 2). Au regard des critères d'évaluation et de sélection des archives, la conservation ou l'élimination de la copie numérique ou de l'original papier dépendra d'un certain nombre de principes, critères et considérations examinés ci-après.

3.2.1. Le principe de l'élimination contrôlée

Si on se base sur la pratique dans l'univers papier, il semble inutile et redondant de conserver à la fois les originaux et ses multiples copies. Au vu de la croissance exponentielle de la production documentaire, il est même impératif de procéder à un tri sélectif afin de ne conserver que l'essence même de l'information.

⁷³ En ce sens, D. Mougenot, « Le régime probatoire de la photocopie et du télécopie », in *La preuve*, Liège, Formation permanente CUP, Liège, 2002, p. 242 et s., n° 10 et s.

⁷⁴ Ainsi, les copies réalisées par les organismes d'allocations familiales (art. 173^{ter} des lois belges coordonnées du 19 décembre 1939 relatives aux allocations familiales pour travailleurs salariés, insérés par la loi du 19 décembre 1990).

⁷⁵ Voy. les textes cités, à l'époque, par D. Mougenot, « Le régime probatoire de la photocopie et du télécopie », in *La preuve*, Liège, Formation permanente CUP, Liège, 2002, p. 242 et s., n° 10 et s. L'auteur constatait déjà un retrait du législateur dans les dispositions adoptées dans les années 90.

⁷⁶ Art. 60 du Code TVA belge et Circulaire n° AFER 16/2008 (E.T.112.081) du 13 mai 2008.

⁷⁷ Art. 6 de la loi belge du 17 juillet 1975 relative à la comptabilité des entreprises.

⁷⁸ Art. 24 et 25 de l'arrêté royal du 8 août 1980 relatif à la tenue des documents sociaux, *M.B.*, 27 août 1980.

⁷⁹ Il s'agit simplement d'une application de la maxime *Lex specialis derogat legi generali*.

⁸⁰ Sur ce problème, voy. également M. Demoulin et S. Soyez, « L'archivage électronique dans le secteur public : entre archivage légal et archivage patrimonial », dans M. Demoulin, *L'archivage électronique et le droit*, Bruxelles, Larcier, 2012, pp. 46-48.

Par analogie, on pourrait donc considérer que la double conservation d'un document dans sa forme originelle papier et dans sa forme copiée numérique est superflue. Toutefois, avant toute destruction de documents, il est indispensable qu'une analyse soit réalisée par les archivistes responsables de l'évaluation et la sélection des archives. Ces derniers sont chargés de contrôler sur le terrain la production documentaire et vérifier si le producteur d'archives ne procède pas à des éliminations inopportunes. Dans le secteur public, il s'agit là du principe dit de l'élimination contrôlée des archives,⁸¹ consacré dans la loi sur les archives : « les autorités ne pourront procéder à la destruction de documents sans avoir obtenu l'autorisation de l'archiviste général du Royaume ou de ses délégués. »⁸² L'idée sous-jacente n'est pas d'interdire la destruction de toute archive, mais de veiller à ce qu'une archive qui présente encore un intérêt scientifique, historique ou social ne soit pas détruite par une administration. Il appartient en effet aux Archives de l'État, et non au service public concerné, d'évaluer si ses archives présentent un tel intérêt et de déterminer leur destination définitive.⁸³

Concrètement, ce principe d'élimination contrôlée se matérialise sous la forme d'un tableau de gestion des documents⁸⁴ dans lequel sont exposées les recommandations précises de conservation et d'élimination d'archives.

3.2.2. Les méthodes d'évaluation

La sélection des archives à conserver est basée sur des méthodes d'évaluation concentrées sur le terrain de production des documents. Ces méthodes reposent sur différents critères, d'ordre juridique, patrimonial et pragmatique.

Sur le plan juridique, en l'état actuel du droit, étant donné la moindre force probante de la copie numérique, on a vu qu'il est risqué de procéder à la destruction du document original papier, lorsque celui-ci destiné à faire preuve en justice. À moins que des dispositions particulières permettent le recours à la copie avec une force probante proche de l'original, les centres d'archives auront tendance à conserver les documents à valeur juridique sous leur format papier, même s'ils ont été numérisés.

Sur le terrain patrimonial, l'archiviste analysera aussi la valeur historique d'une copie numérique par rapport à son original papier. Or, à l'heure actuelle, il apparaît plus aisé d'établir l'authenticité d'un original papier, ce qui plaiderait ici encore pour la conservation de ce dernier, du moins tant que la diplomatique moderne n'a pas atteint une certaine maturité et tant que des procédures homogènes, claires et complètes de numérisation de documents n'ont pas été mises en places et encadrées juridiquement au niveau des producteurs d'archives.

En définitive, les motivations des archivistes à conserver la copie numérique d'un document en plus de son original papier seront pragmatiques. On songe ainsi aux registres papier des actes d'état civil, qui doivent être conservés dans leur forme originale pour des raisons juridiques et historiques, et dans leur forme numérisée, avec leur base de données descriptive, pour un accès en ligne. En effet, lorsque la copie

⁸¹ Pour plus de détails sur ce principe d'élimination contrôlée et sa mise en œuvre en Belgique, voy. M. Demoulin et S. Soyeux, « L'archivage électronique dans le secteur public : entre archivage légal et archivage patrimonial », dans M. Demoulin, *L'archivage électronique et le droit*, Bruxelles, Larcier, 2012, pp. 42-43.

⁸² Art. 5 de la loi belge du 24 juin 1955 sur les archives.

⁸³ Art. 12 de l'arrêté royal sur la surveillance archivistique.

⁸⁴ En terminologie archivistique belge, ce tableau est appelé « tableau de tri des archives »; dans d'autres terminologies : « tableau de conservation », « tableau d'élimination », « tableau de gestion » ou encore « échancier de conservation ».

numérique a été réalisée pour préserver l'original papier de détériorations accidentelles (copie de sécurité) ou pour en faciliter l'accès (copie d'utilisation), la copie numérique présente une utilité en soi et il serait irrationnel de ne pas la conserver sur la seule base de sa moindre valeur juridique.

3.2.3. Conserver l'original papier ou la copie numérique?

À ce stade, on pourrait penser que la question est de savoir si l'on conserve uniquement l'original papier, ou l'original papier avec sa copie numérique. Traditionnellement, le papier jouit en effet d'une certaine préférence dans les milieux juridiques et archivistiques, de sorte que certains pourraient préconiser de le conserver en tout état de cause. On a vu que les critères juridiques et historiques d'évaluation plaideront en ce sens, du moins en l'absence d'un cadre clair et précis sur les procédures de numérisation et la valeur des copies numériques. Les critères pragmatiques conduiront à conserver, en plus du papier, sa copie numérique, comme copie d'utilisation et/ou de sécurité.

En réalité, dès lors qu'une numérisation des documents a lieu, la vraie question qui se pose est de savoir si l'on conserve les deux exemplaires, ou seulement l'un d'entre eux. Dans ce dernier cas, il est piquant de constater que la tendance actuelle semble être de ne conserver que l'exemplaire numérique et de détruire l'original papier, pour des raisons de facilité de gestion et des raisons purement économiques.

Ainsi, lorsqu'une disposition sectorielle octroie une force probante à certaines copies (*supra*, point a), les administrations concernées y voient souvent une autorisation implicite de destruction des originaux papiers. Cette interprétation est d'ailleurs corroborée par la lecture des travaux préparatoires de certains de ces textes : « L'accroissement permanent du volume des informations impose également d'abandonner des archives-papiers difficilement accessibles au profit de copies électroniques ou optiques. (...) Les mesures envisagées permettent en conséquence la destruction des documents originaux et la réduction optimale du problème de l'archivage-papier. »⁸⁵ Mais les travaux préparatoires n'ont pas force de loi. Par ailleurs, l'interprétation qu'ils proposent est contraire à la loi belge sur les archives et outrepassse le champ d'application de ces dispositions dérogatoires, qui ne visent qu'à octroyer une force probante, et non à autoriser la destruction des archives numérisées. Cette situation est problématique lorsque des lots entiers d'archives originales sont détruits sans consultation préalable des Archives de l'État quant à leur éventuelle valeur patrimoniale.

Plus explicitement, on trouve parfois dans la législation même une véritable *obligation de destruction du double* (papier ou électronique), afin d'éviter la gestion en parallèle de documents identiques sur des supports différents. En France, le décret sur la gestion du dossier individuel des agents publics sur support électronique prévoit que « le dossier individuel peut être créé et géré, en tout ou partie, sur support électronique, soit à partir de documents établis sur support-papier et numérisés, soit à partir de documents produits directement sous forme électronique. (...) *En cas de coexistence des supports électronique et papier, toute pièce versée au dossier ne peut être conservée que sur l'un des deux supports, selon le mode de gestion choisi par l'administration.* »⁸⁶ L'idée est que si le dossier est composé

⁸⁵ Voy. le rapport au Roi de l'Arrêté royal du 9 janvier 2000 relatif à la force probante des informations utilisées par l'Administration des Pensions pour l'application de la législation dont elle est chargée et le rapport au Roi de l'Arrêté royal du 15 mars 1999 relatif à la valeur probante, en matière de sécurité sociale et de droit du travail, des informations échangées, communiquées, enregistrées, conservées ou reproduites par les services ministériels et les parastataux du Ministère de l'Emploi et du Travail.

⁸⁶ Art. 2 du décret français n° 2011-675 du 15 juin 2011 relatif au dossier individuel des agents publics et à sa gestion sur support électronique.

de documents nativement électroniques et de documents papier et que l'administration ne souhaite pas travailler avec un dossier hybride, elle a le choix : soit elle opte pour un dossier entièrement papier et procède alors à l'impression des documents électroniques,⁸⁷ soit elle opte pour un dossier électronique et procède à la numérisation des documents papier. Dans un cas comme dans l'autre, elle devrait ensuite détruire l'original, conformément au décret. « Lorsque l'autorité administrative ou territoriale chargée de la gestion du dossier crée une copie sur support électronique d'un acte original établi sur support papier, elle utilise un système de numérisation dans des conditions et sous des formes garantissant sa reproduction à l'identique et la conservation pérenne du document ainsi créé. *La copie conforme ainsi établie se substitue au document original sur support papier qui est détruit* dans un délai fixé par [voie réglementaire]. »⁸⁸

Outre ces obligations ou autorisations de destruction, on constate que, du côté des producteurs d'archives, la tendance actuelle semble aussi s'orienter vers la seule conservation de la copie numérique comme copie de substitution, au détriment du papier. En effet, étant donné les ressources matérielles et humaines de plus en plus limitées allouées à la préservation du patrimoine archivistique, il s'avère souvent indispensable d'opérer un choix, la conservation des deux exemplaires en parallèle étant trop coûteuse. Vu les moyens importants déjà investis dans les projets de numérisation⁸⁹ et les avantages pratiques de la copie numérique, celle-ci sera de plus en plus souvent privilégiée. Mais là aussi, l'intervention de l'archiviste dans le contrôle de ces éliminations nous semble cruciale pour garantir une réelle transparence administrative. L'original papier ne devrait être conservé que s'il présente une valeur juridique ou historique importante et à condition de disposer des moyens financiers suffisants. Pour le reste, il y a fort à parier que les archives papier ordinaires seront détruites après leur numérisation, à plus ou moins bref délai. Il s'agit là d'un véritable dilemme pour l'archiviste et d'un bouleversement dans la pratique archivistique.

4. Synthèse et perspectives : vers une reconnaissance de la copie numérique?

Cet examen du cadre juridique et archivistique régissant le statut de la copie numérique permet de mettre en lumière l'inadéquation et les lacunes du cadre actuel. Cet état des lieux nous conduit à penser qu'une réforme en la matière est non seulement nécessaire, mais presque inévitable.

4.1. L'inadéquation et les lacunes du cadre actuel

Sur le plan strictement conceptuel, la copie numérique ne prête guère à controverse. Il s'agit clairement d'une archive. C'est en outre une copie, aux yeux du droit comme de l'archivistique. Dans certains cas, elle pourra même être qualifiée de copie authentique, au sens juridique du terme, s'il s'agit de la copie numérique d'un acte authentique papier, réalisée dans certaines conditions. Au sens archivistique, un

⁸⁷ On rencontre encore cette option en pratique, mais il n'est guère conseillé de procéder à la destruction des fichiers électroniques originaux dans ce cas. En effet, cela entraînerait la perte de tous les avantages du numérique en termes d'accessibilité et de gestion, ainsi que la perte de l'information périphérique au document électronique (métadonnées, certificat, log, etc.).

⁸⁸ Art. 3 du décret français n° 2011-675 du 15 juin 2011 relatif au dossier individuel des agents publics et à sa gestion sur support électronique.

⁸⁹ En Belgique, les Archives de l'État ont investi plusieurs millions d'euros dans des opérations de numérisation patrimoniale ces dernières années.

service d'archives public pourra également délivrer des copies numériques authentiques et certifiées conformes des archives papiers qu'il détient.

En revanche, sur le plan de sa valeur, le statut de la copie numérique est très incertain. À l'heure actuelle, hormis quelques cas particuliers, elle reste généralement considérée comme inférieure à l'original papier, tant par le droit que par l'archivistique et la diplomatique. En cas de contestation de la conformité de la copie à l'original papier, si ce dernier est détruit et que seule la copie numérique subsiste, sa valeur probante sera, sinon nulle, en tout cas inférieure à celle de l'original, selon l'appréciation du juge. Quant à l'expertise de son authenticité au regard de la diplomatique classique, elle sera difficile à établir en l'absence de l'original.

Or, ces considérations juridiques et archivistiques sont de plus en plus fréquemment éclipsées par des motifs pragmatiques et financiers. Sur le terrain, une pression accrue se fait sentir en faveur d'une gestion documentaire homogène, en procédant à la numérisation des flux papier existants, non seulement pour faciliter leur gestion et leur accessibilité en préservant les originaux papiers des détériorations, mais aussi, à l'inverse, pour détruire les originaux papiers, considérés comme trop volumineux et coûteux à conserver en parallèle. Au vu de ce qui précède, on ne peut que constater un décalage flagrant entre les principes et la pratique.

Outre l'inadéquation des règles actuelles, le problème réside plus spécifiquement dans l'absence d'encadrement et de contrôle des procédures de numérisation, source d'incertitudes et de dérives. Plusieurs conséquences graves en découlent. En amont, on constate que certains projets de numérisation-destruction sont réalisés à moindres coûts, souvent au détriment des règles de l'art et sans concertation avec les Archives de l'État. Outre le fait que des archives précieuses peuvent ainsi disparaître, le résultat d'une mauvaise numérisation peut s'avérer catastrophique, notamment en raison de la piètre qualité de l'image numérique, de l'absence de métadonnées de description ou de gestion, ou du choix de formats difficile à exploiter et à conserver. Cette situation conduira inévitablement à une impasse, les copies numériques étant inutilisables et l'original papier ayant disparu. Par ailleurs, si la numérisation a été bâclée, il sera d'autant plus difficile d'apprécier la valeur probante et diplomatique de la copie numérique. Enfin, même si la numérisation a été réalisée avec soin, l'appréciation ultérieure de sa valeur juridique et diplomatique restera incertaine, en l'absence de critères d'évaluation clairs et prédéfinis.

4.2. Nécessité d'une réforme globale et transversale

Le mouvement de numérisation est en marche et constitue une évolution dont il faut tenir compte. Prôner l'immobilisme et le maintien dogmatique des règles actuelles irait à contre-courant d'une tendance irréversible et ignorerait le problème urgent des numérisations anarchiques. Confiner radicalement la copie numérique dans un statut juridique et diplomatique d'infériorité reviendra à laisser planer un doute difficilement tenable sur des pans entiers de moyens de preuve et d'archives, au détriment de la sécurité juridique et de la vérité historique. Dès lors, on ne saurait trop insister sur la nécessité d'encadrer les pratiques de numérisation de manière conjointe et transversale, tenant compte des enjeux juridiques, archivistiques et technologiques, afin de développer une plus grande confiance à l'égard de ces processus et des documents électroniques qui en sont issus.

Sur le plan juridique, on a vu que les premières initiatives du législateur ont jusqu'ici été timides et sectorielles. Le temps est venu d'offrir à la copie numérique une véritable reconnaissance juridique et un statut équivalent à celui de l'original. Il conviendra cependant de mesurer la portée de la réforme, de définir les conditions fonctionnelles et procédurales d'une telle équivalence et d'encadrer clairement la

possibilité de détruire l'original. Répétons-le, l'élaboration d'un tel cadre devrait s'opérer en concertation avec l'archivistique et la diplomatique modernes, afin de développer des principes cohérents et homogènes.

Sur le plan archivistique, compte tenu de ce nouveau cadre juridique, il sera nécessaire d'organiser les pratiques de numérisation afin que les documents numérisés réunissent les qualités nécessaires à l'établissement de leur authenticité et de leur fidélité à l'original, le tout couplé à un processus de préservation numérique assurant le maintien de ces qualités. Le principe de l'élimination contrôlée des archives devrait être maintenu, mais les critères d'élimination adaptés. Loin de nous l'idée de prôner la destruction massive et systématique du patrimoine archivistique original après sa numérisation. Nous invitons simplement à reconsidérer certaines habitudes avec un regard réaliste, procéder à une juste évaluation des risques et des opportunités pour revoir notre attachement au papier et envisager la possibilité de sa destruction, à condition qu'il laisse une trace numérique fiable, pérenne et exploitable.

Les constatations qui précèdent confirment encore la nécessité de faire évoluer et d'adapter la diplomatique classique au nouvel environnement de production lié au monde numérique. Une nouvelle science est ainsi en train de se développer sous le nom de diplomatique numérique, notamment suite aux études de Luciana Duranti et aux travaux du groupe de recherche InterPARES⁹⁰ ou, en France, de Marie-Anne Chabin. On ne peut que se réjouir de cette évolution, en souhaitant qu'elle continue à se développer de concert avec les juristes.

Naturellement, lors de l'élaboration de ce cadre juridique et archivistique global, il sera indispensable de tenir compte des règles de l'art, bonnes pratiques, normes et standards existants en la matière, tout en veillant à leur adaptation aux nécessités juridiques et archivistiques.

Enfin, ces mesures resteront vaines sans une sensibilisation et une formation des acteurs de terrain (producteurs d'archives, archivistes, diplomates, juristes, informaticiens...) à une numérisation de qualité et une bonne information des contraintes juridiques et patrimoniales en la matière. Les Archives de l'État sont d'ailleurs appelées à jouer un important rôle de conseil à cet égard.

Seule une approche globale et transdisciplinaire permettra d'encadrer la numérisation de manière appropriée et cohérente. Il s'agit là d'un défi à relever de manière concertée par les juristes, les professionnels de la préservation du patrimoine et les informaticiens, dans le souci de préserver le patrimoine archivistique durablement et dans un climat de confiance.

⁹⁰ Cf. www.interpares.org.

Web Archiving as Part of Building the Documentary Heritage of Our Time

Chinese Web Archiving and Statistical Analysis on Chinese Web Archives

Liu Hua,¹ Yang Menghui,² Zhao Guojun and Feng Huiling

School of Information Resources Management, Renmin University, China

¹liuhua2011@ruc.edu.cn, ²yangmenghui@ruc.edu.cn

Abstract

Two Chinese Web archiving projects, Web InfoMall and WICP, are introduced, and two Chinese Web Test collections from Web InfoMall are analysed. The size of Chinese Web pages, Web sites, domains are counted. The top level domain number distribution and the second level domain number distribution in ccTLD(country code Top-Level Domains) ‘.cn’ in the Chinese web pages are analysed. Furthermore, the distributions of sizes of the strongly connected component are analysed. The results show that Chinese Web shows the approximately same nature of the global web, e.g., the domain distribution and connectivity, although there exist some differences in some special properties, e.g., the number of links per page and the number of pages per site are very higher than that of global web. Chinese web archiving initiatives have now begun to move into a more practical implementation phase for the whole Chinese Web sites collecting and long-term preserving.

Authors

Liu Hua is currently a Ph.D. candidate in the School of Information Resource Management, Renmin University of China. She received her Master’s Degree from Beijing Normal University in 2001. Her research interests include Web archiving, Web archives analysis, information resource management, knowledge management. She currently conducts a research project of “knowledge discovery in Oceanic Chinese Web information resources and its application,” which is supported by Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China.

Yang Menghui is currently an associate professor in the School of Information Resource Management, Renmin University of China. He was awarded the Ph.D. degree in Computer Science and Technology from Beijing University of Posts and Telecommunications in 2005. He was a Post-doctor in Department of Computer Science and Technology, Tsinghua University, from Sep. , 2005 to Dec., 2007. He conducts research in the areas of Design, Implementation and Analysis on Communication system and Computer Networks including Internet, Web Mining, Wireless Communication and Mobile Service System.

Zhao Guojun, Professor, Doctor of Management, doctoral supervisor, president of School of Information Resource Management in Renmin University of China, member of Disciplinary Examination and Appraisal Group of the State Council, committee member of the National Electronic Business Standardization Technical Committee, member of Expert Panel of National Civil Servant Employment Examination, committee member of Foundation Theory Committee of the Chinese Archivists Society, advisor of the China Software Parent Company on electronic government.. His research interests include Government information management, electronic government, government agency management, total quality management of non-profit organization, management system design.

Feng Huiling, Professor, Doctor of Management, doctoral supervisor, vice-president of Renmin University of China, vice president of the Society of Chinese Archivists and chair of the Academic Committee of Basic Theory of Archival Science, vice president of the Society of Beijing Archivists, director of the Supervisory Committee of Archival Teaching of the Ministry of Education, corresponding member of the Section for Education and Training of the International Council on Archives (ICA/SAE). Her research interests include Archives management, document retrieval, electronic records management.

1. Introduction

There are a number of reasons why the web can be difficult to collect and preserve. Some of these are technical, e.g., related to the size and nature of the web itself, while others are related to legal or organizational issues.¹ One general problem is that the web is huge and still growing. Table 1.1 is an attempt to show the relative size of selected web archives as of late 2011.

Table 1.1. Approximate size of selected Web archiving initiatives, 2012.²

Initiative	Country	Archived contents (web pages)	Disk space Occupied	Creation Year
Internet Archive	USA	150 billion	5.500 PB	1996
BnF-BnF Web Legal Deposit	France	14 billion	200 TB	2006
Netarkivet.dk	Denmark	7.9 billion	230 TB	2005
PANDORA(Australia's Web Archive)	Australia	3.1 billion	105 TB	1996
Web InfoMall	China	3 billion	100 TB	2001
Sweden (Kulturarw3)	Sweden	1.7 billion	72 TB	1996

Although the figures are approximate and do not take into account things like compression or crawling frequency, it shows that the largest collection is the Internet Archive with over 150 billion web pages and 5.5 PB of data (and growing).

Based on the report of China Internet Network Information Center (CNNIC) in January, 2012, by the end of 2011, there are 7.75 million domains in Chinese Web, the number of Chinese country code Top-Level Domains ‘.cn’ is 3.53 million, and the Chinese Web sites are about 2.30 million.³

Chinese Web archiving is the process of collecting portions of the Chinese Web pages which include Chinese content and ensuring the collection is preserved in an archive, such as an archive site, for future researchers, historians, and the public. Due to the massive size of the Web, Web archivists typically employ Web crawlers for automated collection. Two large Chinese Web archiving projects based on an automatic crawling approach are the Web InfoMall in Peking University⁴ and the WICP (Web Information Collection and Preservation) in National Library of China,⁵ The Web InfoMall is the largest

¹ M. Day, “Preserving the fabric of our lives: a survey of Web preservation initiatives,” in *Proceedings of 7th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2003, Trondheim, Norway, August 17-22, 2003* (Springer-Verlag, 2003), 461-472.

² Wikipedia, “List of Web archiving initiatives,” accessed March 20, 2012, http://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

³ CNNIC, the 29th Statistical Report of Chinese Internet Development, Chinese Internet Network Information Center, accessed January 16, 2012, http://www.cnnic.cn/research/bgxz/tjbg/201201/t20120116_23668.html.

⁴ H. F. Yan and X. Li, “On the Structure of Chinese Web,” *Journal of Computer Research and Development* 39 (2002): 958-967.

⁵ L. Chen, S. Z. Hao, and Z. G. Wang, “Web Information Resource Collection and Preservation—Introduction of WICP and ODBN in the National Library of China (in Chinese),” *Journal of The National Library of China* 1 (2004): 1-6.

and most comprehensive Chinese web archive site for collecting and preserving Chinese Web information in China. It offers permanent storage and access to collections of Chinese Web history information. By the end of 2010 Web InfoMall preserved more than 3 billion Chinese Web pages in more than 8 million domain and 20 million sites from 2001. WICP preserved all the Chinese government web pages (‘.gov.cn’) in more than 80 thousand government sites, E-journal, online newspapers and Chinese Studies from 2003. There are more than 100 special topics, such as 2008 Beijing Olympic Games, SARS, the manned space flight project, etc. The volumes are more than 18TB. In this paper, we have analysed two data sets of Chinese Web archives from Web InfoMall, which were the history Chinese web pages in 2004 and 2006, and get some statistic data on Chinese Web.

2. Chinese Web Archiving Projects

On January, 2002, the first batch of Chinese web pages was archived in Web InfoMall[4]. About 1.5 million pages incremental a day since then. As of today, Web InfoMall has accumulated over 3 billion Chinese web pages, and total online data volume is about 100TB, with an offline backup. It provides access to previous web information, such as fetching a historical Chinese web page, browsing previous web pages. This temporal search is especially useful for studying historical events. Fig.1 shows the architecture of Web InfoMall. The goal of Web InfoMall is to fetch and archive as much Chinese web

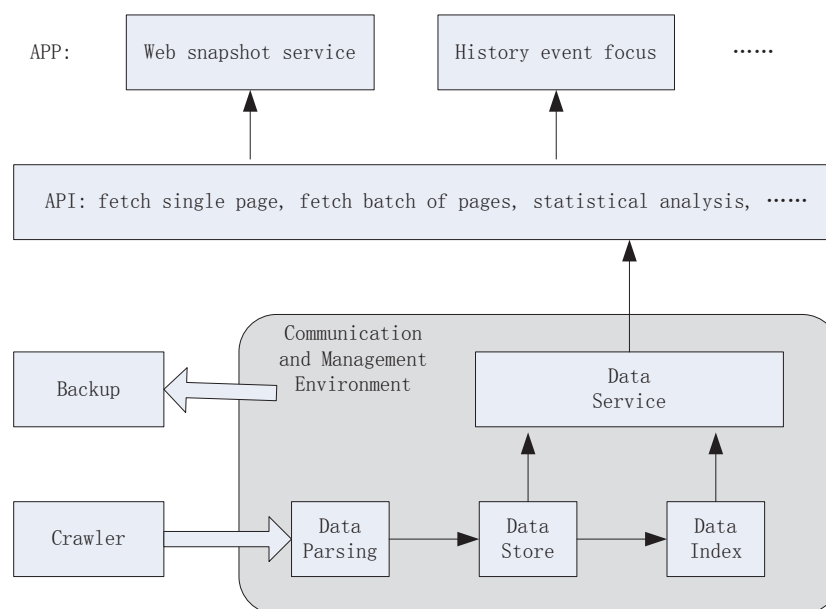


Figure. 1. Web InfoMall in Peking University.

pages as possible before they are gone. The average life cycle of a random web page is about 100 days. Life cycle of ‘.com’ domain web pages is shorter, and that of ‘.gov’ domain web pages is longer. 50% of current viewable web pages will disappear in about 1 year. The number of Chinese Web sites is about 2 million, and the total Chinese Web pages are about 60 billion; The average Web pages per Web site are about 30 thousand. These statistical data show that Chinese Web differs from other country’s Web. Further study on Chinese Web archives by extracting the strongly connected components shows that the

connection between Chinese web pages and Chinese web sites both increases quickly from 2004 to 2006. Today's Chinese Web is a heavily connected network between Chinese web sites.

WICP was established in 2003 to initiate a pilot program to collect and preserve Chinese Web information, and Fig.2 shows the framework of WICP[5]. WICP focus on harvest of the web pages about the events that have great influence on the society, economy and so on , and the sites in 'gov.cn' domain. By the end of 2010, it preserved all the Chinese government web pages('gov.cn' domain) in more than 80 thousand government sites, 315 classifications of E-journal, online newspapers and Chinese Studies from 2003. There are more than 100 special topics, such as 2008 Beijing Olympic Games, SARS, the manned space flight project, etc. The volumes are more than 18TB. On line database navigation can access to about 20 thousand service items, such as government information, library in China and abroad, E-journal, and all special topics, etc.

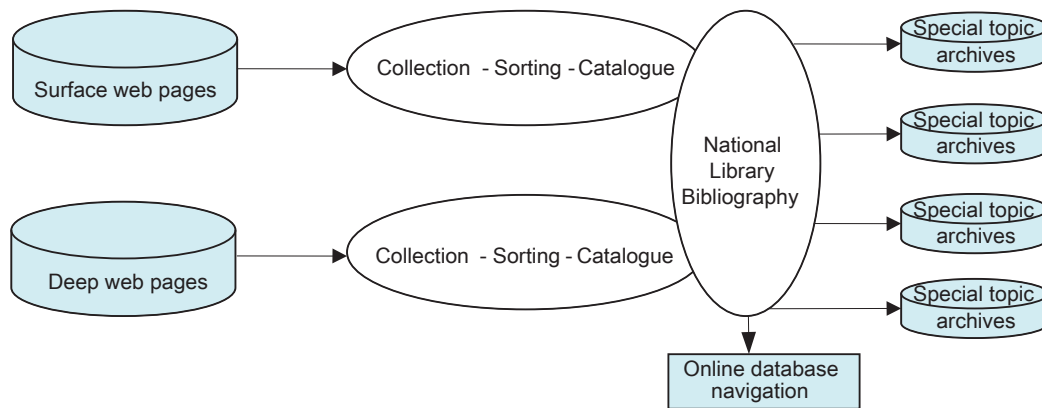


Figure 2. WICP in National Library of China.

3. Analysis on 2 data set from Web InfoMall

Two Chinese Web archives from Web InfoMall are analysed in this paper. Table 3.1 shows the details of these two data sets.

Table 3.1. Chinese Web Test collection, cwt100g and cwt200g.

Data set	Year	Size	Hosts	Pages	Pages/host
Cwt100g	2004	90GB	17045	4737349	278
Cwt200g	2006	197GB	29184	32223476	1104

Cwt100g(Chinese Web Test collection with 100 GB web pages) includes 17045 hosts and 4737349 pages in June, 2004, and the size is about 90GB. In this data set, 69% outlinks in Chinese Web sites link to the same site. In those outlinks which link to other sites, 81% outlinks link to the local province sites. The average pages per site are about 278. Cwt200g(Chinese Web Test collection with 200 GB web pages) includes 29184 hosts and 32223476 pages in April, 2006, and the size is about 197GB. The average pages per host are about 1104.

3.1 Analysis on Web Pages

We consider the Web a hierarchical system in which web pages are the bottom layer, which is operationally defined by a complete URL, such as <http://www1.ruc.edu.cn/101587/25416.html>.

In the experiment, we use two China Web dataset crawled by Peking University Sky Net search engine in June, 2004 and in April, 2006. The size of the raw data is nearly 300G, which contains hyperlinks from the source page to the destination page. Of the 5.6 million pages in cwt100g, 160 million links are found, which amounts to 29 links per page. Of the 37 million pages in cwt200g, 2 billion links are found, which amounts to 54 links per page.

The top level domain number distribution presents differently from the Chinese Web pages. From Table 3.2 we can notice ‘.com’ accounts for most of the domain number, ‘.cn’ next, then ‘.net’, ‘.org’ in cwt100g dataset. Furthermore, we count the second level domains in ccTLD ‘.cn’ in the cwt100g. From Table 3.3 we can find ‘.com’ accounts for most, ‘.org’ next, then ‘.gov’, ‘.edu’, ‘.net’ follow.

Table 3.2. Top Level Domains in the web pages of cwt100g data set.

Top Level Domain	Number	Percentage
.com	2646480	55.8642%
.cn	1545828	32.6307%
.net	430892	9.09564%
.org	53016	1.11911%
.tv	9594	0.202518%
.cc	2859	0.0603502%
.info	1762	0.0371938%
Total	4737349	99 %

Table 3.3. Second Level Domains in ccTLD (country code Top-Level Domains) ‘.cn’ in the web pages of cwt100g.

Second Level Domain	Number	Percentage
.com.cn	556820	36.0208%
.org.cn	415809	26.8988%
.gov.cn	235818	15.2551%
.edu.cn	167119	10.811%
.net.cn	55316	3.57841%
Others	114946	7.43589%

From Table 3.4 we can notice ‘.com’ accounts for most of the domain number, ‘.cn’ next, then ‘.net’, ‘.org’ in cwt200g dataset. Furthermore, we count the second level domains in ccTLD ‘.cn’ in the cwt200g. From Table 3.5 we can find ‘.com’ accounts for most, ‘.gov’ next, then ‘.edu’, ‘.net’, ‘.org’ follow.

Table 3.4 Top Level Domains in the web pages of cwt200g.

Top Level Domains	Number	Percentage
.com	20078413	62.3099%
.cn	7799689	24.205%
.net	3821660	11.8599%
.org	455799	1.41449%
.info	52351	0.162462%
.gov	0	0
.mil	0	0
Total	32223476	99.9%

Table 3.5. Second Level Domains in ccTLD (country code Top-Level Domains) '.cn' in the web pages of cwt200g.

Second Level Domain	Number	Percentage
.com.cn	2724857	34.9355%
.gov.cn	511784	6.56159%
.edu.cn	372310	4.7734%
.net.cn	280094	3.59109%
.org.cn	172068	2.20609%
Others	3738576	47.9324%

3.2 Analysis on Web Sites

Host is considered the second layer, which is defined as the collection of web pages hosted on a web server. More precisely, a host corresponds to all the pages under the address of `http://.../` (i.e., the part between “`http://`” and the first “`/`” from the left). Of the 5.6 million pages in cwt100g, 17045 hosts are found, which amounts to 278 pages per host. Of the 37 million pages in cwt200g, 29184 hosts are found, which amounts to 1104 pages per host.

The top level domain number distribution presents differently from the Chinese Web sites. From Table 3.6 we can notice ‘.com’ accounts for most of the domain number, ‘.cn’ next, then ‘.net’ and ‘.org’ in cwt100g dataset. Furthermore, we count the second level domains in ccTLD ‘.cn’ in the cwt100g. From Table 3.7 we can find ‘.com’ accounts for most, ‘.gov’ next, then ‘.edu’, ‘.net’, ‘.org’ follow.

From Table 3.8 we can notice ‘.com’ accounts for most of the domain number, ‘.cn’ next, then ‘.net’ and ‘.org’ in cwt200g dataset. Furthermore, we count the second level domains in ccTLD ‘.cn’ in the cwt200g. From Table 3.9 we can find ‘.com’ accounts for most, ‘.gov’ next, then ‘.edu’, ‘.net’, ‘.org’ follow.

The statistic data indicates the development of World-Wide Web in China is not in balance, the Web is used mostly for commerce in China. It is also true for global web. According to the top level domain distribution, ‘.com’ accounts for most, and ‘.cn’ is the second most. According to the top level domain distribution in ccTLD ‘.cn’, ‘.com.cn’ accounts for most, and ‘.gov.cn’ is the second most.

Table 3.6. Top Level Domains in the web sites of cwt100g data set.

Top Level Domains	Number	Percentage
.com	10886	63.8662%
.cn	4032	23.655%
.net	1850	10.8536%
.org	277	1.62511%
.edu	0	0
.gov	0	0
.mil	0	0
Total	17045	100%

Table 3.7. Second Level Domains in ccTLD (country code Top-Level Domains) '.cn' in the web sites of cwt100g.

Second Level Domain	Number	Percentage
.com.cn	2186	54.2163%
.gov.cn	704	17.4603%
.edu.cn	398	9.87103%
.net.cn	201	4.98512%
.org.cn	117	2.90179%
Others	426	10.5655%

Table 3.8. Top Level Domains in the web sites of cwt200g.

Top Level Domain	Number	Percentage
.com	19132	65.5565%
.cn	6987	23.9412%
.net	2621	8.98095%
.org	323	1.10677%
.edu	0	0
.gov	0	0
.mil	0	0
Total	29184	99.6%

Table 3.9. Second Level Domains in ccTLD(country code Top-Level Domains) '.cn' in the web sites of cwt200g

Second Level Domain	Number	Percentage
.com.cn	1993	28.5244%
.gov.cn	832	11.9078%
.edu.cn	821	11.7504%
.net.cn	204	2.91971%
.org.cn	185	2.64777%
Others	2952	42.2499%

3.3 Analysis on connectivity

Power law distributions have been observed in various aspects of the web.⁶ Large lots of pages with small in-degrees (or out-degrees) coexist with small but not negligible pages with large in-degrees (or out-degrees). This phenomenon reflects the unbalanced nature of Web links. We show that power laws also arise in the distribution of sizes of the strongly connected components (SCCs). By running the strongly connected component algorithm, we find that there are 17867909 SCCs in cwt100g, and there is a single large SCC consisting of about 114137 pages. Fig. 3 shows the distribution of sizes of the strongly connected components in cwt100g data set.

There are 930375074 SCCs in cwt200g, and there is a single large SCC consisting of about 3486115 pages, all other components are significantly smaller in size. This amounts to barely 11% of all the pages in cwt200g. Fig. 4 shows the distribution of sizes of the strongly connected components in cwt200g data set.

4. Long-term Preservation

Many current web archiving initiatives have been focused on the collection of web resource, but there remains a need to consider how those web sites being collected at the moment can be preserved over time, and what this may mean. This may include assessment of various proposed preservation strategies and implementation of repositories based.

Web InfoMall uses a customized storage format and standard to preserve web pages.⁷ It uses the Google File System,⁸ Chubby,⁹ and Bigtable¹⁰ technologies to implement a distributed storage system for history Chinese Web pages. Based on the storage system, it provides Wayback Machine to access on the old Chinese Web pages. Three access types are supported: (1) Users know the exact URLs that they are looking for, and users provide the date that they want to see what the old Web pages looked like. Such as access to the <http://www.pku.edu.cn/index.html> in September 1, 2008. Of course the Web page in this time point in the system may not be stored, naturally it will return the recently Web page version before this time. For example, before September 1, 2008, the system stored the recently Web page version in August 15, 2008. It will return this page data. (2) Users know the exact URLs that they are looking for. Such as access to the web version of <http://net.pku.edu.cn/index.html>. It will return all versions of this web page. (3) Users want to access to a certain range of URL webpage snapshot. If the requested access conforms to *.pku.edu.cn domain name in July, 2008. The range of URL must be a continuous URL space, it returns the content in accordance with the URL character sequence from small to larger order.

⁶ A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, S. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Computer Networks* 33 (2000): 309-320.

⁷ Lianen Huang, "On the Technologies for Building and Accessing a Web Archive," Ph.D diss., Peking University (in Chinese), 2008, pp. 112-114.

⁸ S. Ghemawat, H. Gobioff, and S. T. Leung, "The Google file system," *ACM SIGOPS Operating Systems Review* 37 (2003): 29-43.

⁹ M. Burrows, "The Chubby lock service for loosely-coupled distributed systems," in *OSDI '06: 7th USENIX Symposium on Operating Systems Design and Implementation*, 2006, pp. 335-350.

¹⁰ F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," in *OSDI '06: 7th USENIX Symposium on Operating Systems Design and Implementation*, 2006.

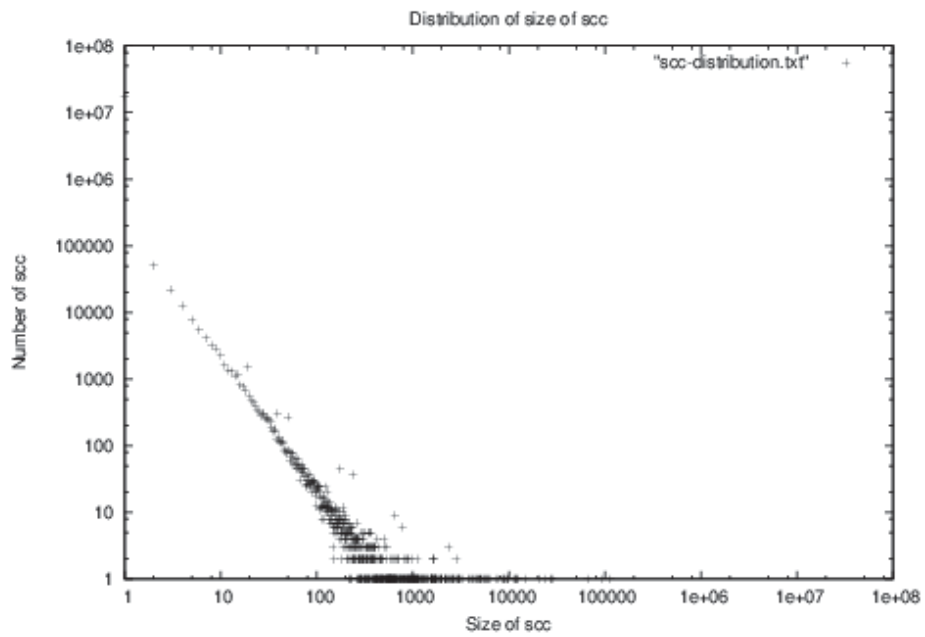


Figure 3. SCC (strongly connected components) distribution in cwt100g data set.

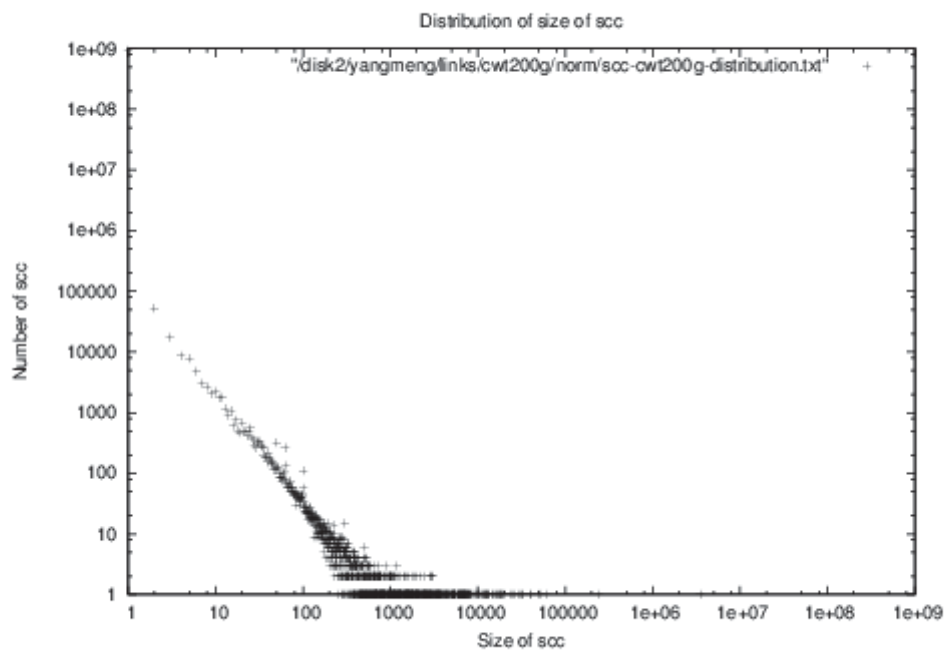


Figure 4. SCC (strongly connected components) distribution in cwt200g data set.

Web information preservation at National Library of China uses Open Archival Information System reference model,¹¹ and adopts Machine-Readable catalogue and Dublin Core metadata standard sets. Technological attempts of digital information preservation, such as reformatting and migration, are also used in this project. Although there has not come up with an effective way to preserve digital resources, WICP is ready to work with all colleagues in library community to preserve digital information.

5. Conclusion

Chinese Web archiving initiatives has demonstrated that collecting and preserving Web pages is an interesting area of research and development that has now begun to move into a more practical implementation phase. This paper introduces two Chinese Web archiving projects, Web InfoMall and WICP. Two data sets of Chinese Web archives from Web InfoMall are analysed. The results show that Chinese web has the same nature of the global web, e.g., the domain distribution and connectivity, and there also exists some differences in some special properties, e.g., the number of links per page and the number of pages per sites are very higher than that of global web.

Acknowledgements

This paper is supported by Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (No.12XNH181).

¹¹ CCSDS 650.0-B-1: *Reference Model for an Open Archival Information System (OAIS)*, Consultative Committee on Space Data Systems, January 2002, accessed March 20, 2012, <http://public.ccsds.org/publications/archive/650x0b1.pdf>.

Implications of the Web Semantization on the Development of Digital Heritage

Gustavo Urbano Navarro

*Universidad Nacional de la Patagonia Austral. Unidad Académica San Julián, Argentina,
gusunavarro@gmail.com*

Abstract

The general objective of this document lies in the construction of a network-interoperable digital repository (DR), of open access for the filing, preservation and distribution of documentary material- local press and photographs of the first half of XX century concerning towns in the Patagonia region (Puerto Deseado, Puerto Santa Cruz, Comandante Luis Piedrabuena y Puerto San Julián) -, for the materialization of its available use within the academic environment, as well as in the secondary and tertiary levels of the educational system, and the community in general. Its result will consist of developing an experimental prototype focused on technical, descriptive and structural processes of the DR (metadata schema), and its interoperability.

Author

Dr. Gustavo Urbano Navarro is Director of proyect Archivo Memoria de la Patagonia Austral. Koluel. Ph.D. Universidad de Las Palmas de Gran Canaria. Centro de Innovación para la Sociedad de la Información. CICEI and Profesor Universidad Nacional de la Patagonia Austral. Unidad Aacadémica San Julián. Santa Cruz. Argentina and Universidad Nacional de San Martin. UNSAM. Buenos Aires. Argentina.

1. Memory of the Patagonia Austral¹

The general objective of this document lies in the construction of a network-interoperable digital repository (DR), of open access for the filing, preservation and distribution of documentary material- local press and photographs of the first half of XX century concerning towns in the Patagonia region (Puerto Deseado, Puerto Santa Cruz, Comandante Luis Piedrabuena y Puerto San Julián), for the materialization of its available use within the academic environment, as well as in the secondary and tertiary levels of the educational system, and the community in general. Its result will consist of developing an experimental prototype focused on technical, descriptive and structural processes of the DR (metadata schema), and its interoperability.

Thus, the users are intended to retrieve the stored material, by browsing the DR content through the search interfaces. The problem aroused in the Patagonia region responds to the absence of a sustained policy of preservation of the historical heritage at long-term.

As a result of this, the researchers do not have, in many cases, the possibility of access to collections of data or sources existing in municipal libraries, records and local documentation centers, since the contributions coming from the interior regions to the national history are increasingly significant. In our historiography there are lots of generalizing works supported by empirical studies carried out in and from the central area, without devoting too much study to realities and dynamics.

¹ www.koluel.org

Therefore, the proposal is referred to the development of such repository taking into account the specific context of the historical heritage and the scientific legacy of the Patagonia, the preservation and management of digital records, the data guardianship, as well as the widespread and maximization of visibility, use and impact of the regional scientific output in the international community.

That is to say, it has a double objective: on one hand, to preserve those contents and to enable the access and use either for teaching purpose or for academic research, encouraging the specificity on the local aspect; and, on the other hand, to provide the feedback of such research and give support to the electronic publications in the area of Social Sciences in the institutions of the region. Consequently, the creation of this repository intends to retrieve in digital files the regional cultural heritage, aiming to increase its visibility and impact.

2. Introduction

2.1 Research and Development Lines

This paper is structured around two convergent axes:

1. The digital preservation of the social historical memory and the cultural heritage of the region: management of digital files for the access at long-term and data curation;
2. The development of an experimental prototype focused in technical processes and in its interoperability with other technological platforms.

3. Achieved/Expected Results

In recent years, information and communication technologies have provided new resources for the management of information, offering additional services for using documentary materials, apart from consultation and organization through the catalogues that digital libraries made available, enlarging and renewing their services, and opening their domain to other types of information resources. Hence, the creation of e-infrastructure appears as a specific solution which consents to collect, get access to and share educational resources, having at disposal a content storage system that is easily integrated and communicated with other operative systems (McLean & Lynch, 2003).

For this reason, e-infrastructures started to take a stand as important tools the function of which is to protect the resources, making them available to be shared with other applications, thus facilitating the content stream. This means that they include apart from storage systems, toolsets which are useful for the reusing process. Although there are portals of the content administrator type, or collaborative work environment, which could certainly meet some of the requirements offered by repositories, from the technological point of view, one of the core problems lies in the fact that the export and import of bibliographic data is usually limited in those cases. In this way, they configure new and complex systems that enlarge the abilities of resources turning them into tools for learning and research.

Municipal libraries and documentation centers of the different towns in the Patagonia region have documentary materials, many of which are getting deteriorated on account of the lack of financing for the maintenance and sustainability of projects for their preservation.

The main objective of the project is to give a possible answer to these problems, through the creation of a new structure for the processing of information thus enabling the data retrieval and contextualization and its integration into flexible networks of knowledge, as well as with social networks.

4. Background

In the early 90s, the first initiatives of open e-infrastructure of specialized documents appeared with the scope of facilitating the access to content, which was restricted so far to those who could afford them. One of the main drivers of the creation of these digital repositories was the Open Access movement in the United Kingdom. The free availability of its content means that any user may read the documents therein contained, without requiring either subscription or registration, that is to say that its availability is free, and the only restriction upon the distribution and reproduction is to grant authors the control on the integrity of their work and the right to be properly quoted and recognized (Budapest Open Access Initiative, 2002).

By the end of 2008, the European Commission launched the “Europeana” project, an e-infrastructure that does not constrain its collection to texts, but includes whichever kind of cultural objects, such as: texts, photographs, videos, maps, manuscripts, paintings, newspapers and historical documents on record, i.e., the contribution of the repositories all over Europe.

The creation of this digital object repository comes up as a solution to facilitate the compilation, the preservation, and the diffusion of differently supported documentary sources, that is it does not limit its collection to texts, but it also includes any kind of cultural objects, such as: texts, photographs, videos, maps, manuscripts, paintings, newspapers and historical documents on record and oral file of voices, providing a content storage system which is integrated and communicated with other systems, enabling the stream of knowledge; and carrying out the following intersectional objectives:

- Protect and foster heritage and cultural expressions through the effective implementation of the Conventions of 1972, 2003, 1954, 1970, 2001; and 2005
- Strengthen the contribution of culture to the sustainable development
- Promote the role of culture in the interaction and dialogue among different cultures in the development policies aiming to social coherence and reconciliation.
- Improve the universal access to information and knowledge

5. Development

During 2011, the team started to think about developing an electronic platform based on a specific software. In order to define this platform, some documents that gave details of the comparison⁴ among Fedora, DSpace and Eprints, were previously read.

We threw ourselves directly to Fedora (Flexible Extensible Digital Object Repository Architecture), which was originally developed by researchers of Cornell University as an architecture to store, to administrate and to accede to digital content under the shape of digital objects inspired in the framework of Kahn and Wilensky (Kahn and Wilensky Framework). The project called Fedora Repository Project implements Fedora’s abstractions in a robust open-source software. It provides central repository services under the shape of web services with well-defined APIs. Moreover, it provides a collection of services and applications such as search, OAI-PMH, messaging, customers and more.

But the problem of Fedora is that its interface was not so friendly; consequently, some development including Fedora was thought to be used, and so Islandora was decided to be tested.

Islandora is an open source project underway at the Robertson Library at eSciDoc fedora the University of Prince Edward Island. Islandora combines the Drupal and Fedora software applications to create a robust digital asset management system that can be used for any requirement where collaboration and digital data stewardship, for the short and long-term, are critical.

There are a number of initiatives under the umbrella of the Islandora project, all developed using the innovative Islandora software suite, including:

- VRE (Virtual Research Environment) - a collection of customized Islandora sites used by researchers at UPEI and elsewhere to steward research data. More information about our VRE's is available from the links on the project's web site.
- IslandArchives.ca - a number of digital collections created by UPEI and partners, providing access to a wide range of digital documents and materials.
- Repository-In-A-Box - a unique solution for building Institutional Repositories and as seen with UPEI's own IslandScholar site.

Especially taking into account that Islandora project provided for the framing of repositories on newspapers, images and voices, different tests were carried out by virtue of Koluel's objectives. After several tests, the decision was not to use Islandora since previous knowledge on Drupal was required and this fact generated a great dependence with Islandora group. Once the testing was finished, it was decided to take much of 2011 in testing other products under Fedora interface. Namely Hydra.

Hydra is a repository solution that is being used by institutions to provide access to their digital content. Hydra provides a versatile and feature-rich environment for end-users and repository administrators alike. The project gives like-minded institutions a mechanism to combine their individual repository development efforts into a collective solution with breadth and depth that exceeds the capacity of any single institution to create, maintain or enhance on its own. The motto of the project's partners is "if you want to go fast, go alone. If you want to go far, go together." Hydra is an ecosystem of components that lets institutions deploy robust and durable digital repositories (the body) supporting multiple "heads": fully-featured digital asset management applications and tailored workflows. Its principle platforms are the Fedora Commons repository software, Solr, Ruby on Rails and Blacklight. And above all, it works in a free software. Unlike Islandora, Hydra is a great framework and it is being developed.

After being performed all tests with Hydra, then the next step was to work with eSciDoc.

- eSciDoc is a platform that contains Fedora-commons inside its structure and contains several applications.
- eSciDoc is a system targeted at research organizations, universities, institutes, and companies interested in eScience-aware knowledge and information management.
- eSciDoc enables you to publish, visualize, manage, and work with data artefacts (or objects). Objects include both publication data and research data across disciplines.
- eSciDoc provides a generic infrastructure and specialized solutions within the context of research questions. It integrates existing solutions and implement new ones.

- eSciDoc addresses aspects of data reliability, data quality, data curation, and long-term preservation. It covers the whole lifecycle of objects.
- eSciDoc supports semantic relations between objects.

The eSciDoc system is designed as a service-oriented architecture (SOA) implementing a scalable, reusable, and 4extensible service infrastructure. Application and discipline-specific applications can then be built on top of this infrastructure. The heterogeneity of the envisioned solutions in addition imposes an efficient handling of different kinds of content.

The service-oriented architecture fosters the reuse of the existing services. An eSciDoc service may be reused by other projects and institutions, either remotely or locally, thus becoming one building block with a broader e-Science infrastructure. At the same time, the SOA approach of eSciDoc comes with other advantages.

Instead of a complex and monolithic application, the eSciDoc service infrastructure is rather to be seen as a set of loosely coupled services, which can be specified and implemented independently. This allows for an iterative implementation strategy for services. First services may already be implemented while others are still in their design phase. Based on feedback from early adaptor users (“pilots”), new services can be easily added, thus fulfilling user expectations in a more timely and user-driven manner.

The core technology used to implement the services is based on Java and XML. Instead of building the infrastructure “from a scratch”, the eSciDoc team chose to integrate existing open-source components as much as possible. eSciDoc services in general provide both SOAP and REST style interfaces. This allows for further development of solutions without constraining the selection of the programming languages, thus accelerating their implementation and enabling the involvement of various developer groups. Even simple scripting and “Web 2.0”-style mash-ups are supported.

The eSciDoc service infrastructure groups its services into three service layers: basic services, intermediate services and application services. The software used under eSciDoc platform was PubMan. PubMan supports research organizations in the management, dissemination and re-use of publications and supplementary material. It can be used out-of-the-box within the eSciDoc infrastructure. Being open source, it will be customized and extended for institutional- and discipline-specific needs. Finally, a last test was carried out with a Content Management System CMS and Plone was chosen.

Plone lets non-technical people create and maintain information for a public website or an intranet using only a web browser. Plone is easy to understand and use—allowing users to be productive in just half an hour—yet offers a wealth of community-developed add-ons and extensibility to keep meeting your needs for years to come. Blending the creativity and speed of open source with a technologically advanced Python back-end, Plone offers superior security without sacrificing power or extensibility.

The Plone community is an incredibly diverse group that bridges many types and sizes of organizations, many countries and languages, and everything from technical novices to hardcore programmers. Out of that diversity comes an attention to detail in code, function, user interface and ease of use that makes Plone one of the top 2% of open source projects worldwide.²

Plone’s intellectual property and trademarks are protected by the non-profit Plone Foundation. This means that Plone’s future is not in the hands of any one person or company.

Thousands of websites across large and small businesses, education, government, non-profits, sciences, media and publishing are using Plone to solve their content management needs. This is

² Source: Ohloh, <http://www.ohloh.net/>.

supported by a global network of over 300 solution providers over 300 solution providers in more than 50 countries. Looking for a hosting provider to host your Plone site for you? You can find a list of providers and consultants on plone.net.

We are very proud to be known by the company we keep. Organizations as diverse as NASA, Oxfam, Amnesty International, Nokia, eBay, Novell, the State Universities of Pennsylvania and Utah, as well as the Brazilian and New Zealand governments— all use Plone.

Plone is open on many levels. It runs on Linux, Windows, Mac OS X, FreeBSD and Solaris—and offers a straightforward installation to get you up and running in minutes. It has been translated into more than 40 languages, and is developed with an unflinching emphasis on usability and standards compliance.

Need a CMS that integrates with Active Directory, Salesforce, LDAP, SQL, Web Services. LDAP or Oracle? Plone does. Need to be sure your website is accessible? Plone meets or exceeds US Government 508 and W3C's WAI-AA standards.

Worried about security? As an open source product, a large number of developers frequently scrutinize the code for any potential security issues. This proactive approach is better than the wait-and-see approach in proprietary software that relies on keeping security issues a secret instead of resolving them outright.

Based on Python and the Zope libraries, Plone has a technological edge that has helped it attain the best security track record of any major CMS.³ In fact, security is a major reason why many CMS users are switching to Plone.

The market is full of open source content management systems, so it is important to do your homework before choosing one for your organization. Remember that a simple CMS may work out great to start with, but lead to problems with scaling or migration when you need more capability than it can provide. At the other end of the spectrum, a powerful CMS can be so difficult to learn and maintain that it never gains acceptance to users. Make sure the CMS you choose meets your needs today without compromising future growth.

In the end, Plone was decided to be used for Koluel project. The decision was taken on account of its friendly interface, its translation into Spanish language among many of the available languages, the integration with OAI and the possibility to program interfaces for cataloguing and classifying documents under Dexterity. Dexterity was created to serve two audiences: Administrators/integrators, and developers.

For administrators and integrators, Dexterity offers:

- The ability to create new content types through-the-web
- The ability to switch on/off various aspects (called “behaviors”) on a per-type basis
- Improved collaboration between integrators (who may define a type's schema, say) and programmers (who may provide re-usable behaviors that the administrator can plug in).

For developers, Dexterity promises:

- The ability to create content types more quickly and easily, and with less boilerplate and repetition, than what is possible with Archetypes or plain CMF types
- Content objects with a smaller runtime footprint, to improve performance
- Types that use the now-standard `zope.interface/zope.schema` style of schema, and more broadly support modern idioms that sit a little awkwardly with Archetypes and its ilk.

³ Source: CVE, <http://cve.mitre.org/>.

As far as koluel is concerned, work was accomplished jointly with the Zest software company, especially with Maurits van Rees in order to develop a product which uses all Dublin Core fields. The product is available in github and can be freely used.

References

- Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, 2003. Accessed 25 July 2010. <http://oa.mpg.de/openaccess-berlin/berlindeclaration.html>.
- Budapest Open Access Initiative, 2002. Accessed 25 July 2010. <http://www.soros.org/openaccess/read.shtml>.
- Bustos González, Atilio, y Antonio Fernández Porcel. "Directrices para la creación de repositorios institucionales en universidades e instituciones de educación superior," 2008. Accessed 12 July 2010. http://www.sisbi.uba.ar/institucional/proyectos/internacionales/Directrices_RI_Espa_ol.pdf.
- Brown, Alejandra M, M. Alejandra González, y Marisa G. Velazco Aldao. s.f. Proyecto piloto de desarrollo de la Biblioteca Digital de Tesis y Disertaciones del Instituto Balseiro. (Inédito)
- De Volder, C. "Los repositorios de acceso abierto en la Argentina. Situación actual." *Información, Cultura y Sociedad* 19 (2008): 79-98. Universidad de Buenos Aires. Facultad de Filosofía y Letras. Instituto de Investigaciones Bibliotecológicas (INIBI), ISSN: 1514-8327.
- Fushimi, Marcela, y Josefina Mallo. "Memoria académica y científica: el rol de la biblioteca universitaria en la preservación y difusión del conocimiento generado en las universidades." Trabajo presentado en las Cuartas Jornadas de Sociología de la UNLP, realizadas en La Plata del 23 al 25 de noviembre de 2005. Accessed 9 July 2010. http://eprints.rclis.org/archive/00010464/01/socio_memoaca_2005.pdf.
- Harnad, Stevan, y Tim Brody. "Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals." *D-Lib Magazine* 10, no. 6. (2006). Accessed 11 July 2010. <http://www.dlib.org/dlib/june04/harnad/06harnad.html>.
- Houghton, J.W., B. Rasmussen, P.J. Sheehan, C. Oppenheim, A. Morris, C. Creaser, H. Greenwood, M. Summers, and A. Gourlay. "Economic Implications of Alternative Scholarly Publishing Models: Exploring the Costs and Benefits." Report to The Joint Information Systems Committee (JISC) by Victoria University & Loughborough University, 2009. <http://hdl.handle.net/2134/4137>.
- Iannello, Catalina, y Inés García Uranga. "Acerca de la Biblioteca Virtual en Salud." *Archivos Argentinos de Pediatría* 106, no. 2 (2008): 97-98.
- Lawrence, Steve. "Free online availability substantially increases a paper's impact." *Nature Web-debates*, 2001. <http://www.nature.com/nature/debates/eaccess/Articles/lawrence.htm>.
- _____. "Online or Invisible?" *Nature* 411, no. 6837 (2001): 521. Accessed 22 July 2010. <http://www.neci.nec.com/~lawrence/papers/online-nature01/>.
- McLean, N., and C. Lynch. "Interoperability between information and learning environments: Bridging the gaps." IMS Global, 2003.
- Melero, Remedios. "Acceso abierto a las publicaciones científicas: definición, recursos, copyright e impacto." *El Profesional de la Información* 15, no. 4 (2005): 255-66. Accessed 12 July 2010. <http://eprints.rclis.org/archive/00004371/01/EPirmelero.pdf>.

- Open Society Institute. *A guide to institutional repository software*, 3rd ed., 2004. Accessed 1 July 2010. http://www.soros.org/openaccess/software/OSI_Guide_to_Institutional_Repository_Software_v3.htmRegistry of Open Access Repositories, <http://roar.eprints.org/>.
- Sánchez, Salvador y Remedios Melero. “La denominación y el contenido de los Repositorios Institucionales en Acceso Abierto: base teórica para la “Ruta Verde.”” 2006. Accessed 18 July 2010. http://eprints.rclis.org/archive/00006368/01/DenominacionB3n_contenido_OA.pdf.
- Sánchez Villegas, Miguel Ángel. “Iniciativas de acceso abierto y perspectivas de E-LIS en México.” Trabajo presentado al *Congreso Internacional de Información INFO 2006*, realizado en La habana, del 17 al 21 de abril de 2006. Accessed 12 July 2010. http://eprints.rclis.org/archive/00006212/01/E-LIS_M%C3%A9xico.pdfhttp://eprints.rclis.org/archive/00006212/01/E-LIS_M%C3%A9xico.pdf.
- SciELO Argentina, Scientific Electronic Library Online. Accessed 28 July 2010. <http://www.scielo.org.ar/scielo.php>.
- Serrano Muñoz, Jordi, y Jordi Prats Prat. “Repertorios abiertos: el libre acceso a los contenidos,” *Revista de Universidad y Sociedad del Conocimiento* 2, no. 2 (2005). Accessed 28 July 2010. <http://www.uoc.edu/rusc/2/2/dt/esp/serrano.pdf>.
- Swan, A., P. Needham, S. Proberts, A. Muir, C. Oppenheim, A. O’Brien, R. Hardy, F. Rowland, and S. Brown. “Developing a model for e-prints and open access journal content in UK further and higher education.” *Learned Publishing* 18, no. 1 (2005): 25-40. Accessed 28 July 2010. <http://eprints.ecs.soton.ac.uk/11000/http://eprints.ecs.soton.ac.uk/11000/>.

Preserving the Web Archive for Future Generations

Practical Experiments with Emulation and Migration Technologies

Matt Holden

Institute national de l'audiovisuel (INA), France, mholden@ina.fr

Abstract

Ina shares the responsibility of the web legal repository in France with the National Library of France (BnF)¹ and has been archiving regularly since the beginning of February 2009. In this paper we look at how we can use emulation/migration technologies to help preserve an obsolete format and as a result, we hope to be able to shed light on which parts of the web archive may be most at risk and whether this can help shape and prioritize our long-term preservation strategy.

Author

Matt Holden, MSci, MSc, ARCS studied for a bachelor's degree in Physics at Imperial College, London before taking a master's degree in Telecommunications at University College London (UCL). This led to an Operations IT role working at Nortel Networks UK in a group developing solutions for the mobile phone GSM-HLR network. Late in 2006, Matt became Data Centre Manager at CMC Markets UK plc, which culminated in a multi-million pound construction project to create a new data centre in East London. In early 2009, he joined Ina with the mission of running IT Operations for the web archiving (DLWeb) team.

1. Introduction

The emulation vs. migration debate is one which has been discussed at great length in many papers and journals too numerous to catalogue. The intention of this paper is to look at practical experiments that can be conducted during the operation of a working web archive.

As Ina is a member of the International Internet Preservation Consortium (IIPC), we were very interested in building upon the work already undertaken by the National Library of Australia (Long, 2009) as part of a project for the IIPC's Preservation Working Group (PWG). We wanted to revisit the work to see how it could be applied to the operation of our own web archive as well as investigate if the tools had changed.

We were also interested in using the process of Transparent Format Migration (Rosenthal et al., 2005) conducted at the LOCKSS (Lots of Copies Keep Stuff Safe) web archive in which image formats were converted automatically when pages were loaded.

Finally, we were curious to find out whether there really were file formats in the oldest part of our archive that were in danger of becoming obsolete.

1.1 The History of Legal Deposit at Ina

Since its creation in 1974 by law, Ina has been in charge of collecting, preserving and making available French audiovisual collections. Initially organized to meet professional needs for a public broadcast

¹ Title IV of French Law n°2006-961 dated 1st August 2006 titled "Loi DADVSI".

archive facility, it quickly grew into a national repository for the French audiovisual heritage. In 1992 due to its existing obligations, Ina was designated by law as being responsible for the Legal Deposit of radio and television. Today, over twenty years later, the institute is considered to be the world's largest broadcast archive, holding over 4-6 million hours of television and radio recordings, dating back to the dawn of the medium. Annually, we see an increase of a million hours of programs from 100 television channels and 20 radio stations which are digitally recorded around the clock.

Following the fast pace of publishing technologies, French Legal Deposit Law was then amended to include the Web, splitting responsibility between the French national library (BnF) and Ina, thus ensuring coherence and continuity of their respective collections. Ina was thus designated as national repository for audiovisual related Web sites—with a broad remit—as well as on-demand audiovisual media services available from Web platforms.

1.2 The History of Web archiving at Ina

Web archiving duly commenced in February 2009 with 3600 seed websites covered in the initial collection process. In late 2010 we worked in partnership with the Internet Archive to recover websites pertaining to our collection from their archive from 1996-2009 (18Tb of data containing 524 million URLs). As of late July 2012, the whole collection contains 15 billion URLs consisting of nearly 10,000 seed websites represented by over 150Tb of compressed data.

1.3 Ina's Web Archive: technologies and statistics

- No URLs are changed during the consultation of the archive
- All access to the archive is controlled via a proxy which matches urls to archived content. We currently use Firefox v3.5 with a number of plugins to facilitate archive access, such as displaying the date of capture and navigation between different archived versions of the same page.
- Ina uses its own Digital Archive File Format (DAFF)
- All content is de-duplicated—there are no multiple copies of objects in the archive, but the objects themselves may be referenced multiple times.²
- When we talk about the overall size of the archive we include multiply referenced objects as this indicates the enormous savings due to the de-duplication process (since sites have been crawled with a high frequency). However from a preservation perspective, it is useful to consider the number of unique objects in the archive (Table 1)

Given the importance of images and sound at Ina, we have concentrated on rich multimedia websites and this is very much reflected in our archived content, with nearly 50Tb of compressed audiovisual material counting for approximately a third of the size of the archive.

² E.g., a 2Mb audio file is captured on a web page five consecutive times whilst a text file of 0.1Mb referencing it changes each time. We have five versions of the text but one single audio file referenced five times, so technically our archive size indicates 10.5Mb as opposed to the 2.5Mb of actual data linked to these versions.

Table 1. Ina web archive statistics (June 2012).

Object Type	Unique objects (millions)	Unique objects size (Tb)	Multiply referenced objects (millions)	Multiply referenced objects size (Tb)
Image	145	3.1	7454	114
Text	1759	82	3390	99
Audio	2.6	25	82	649
Video	1.4	22	29	527
Other	288	4.8	1785	65
Total	2196	137	12740	1364

1.4 Long-term bit preservation (using short-term migrations)

The long-term bit preservation of the archived files falls into line with the typical working processes of many other institutions, using short to medium term migrations (around 20 year strategies). Two copies of the archive are stored on differing generations of disk storage and 2 offline backup copies will be stored on tape. This is based on migrating the archived files onto a new disk storage media every 3-5 years and a new tape based media every 4-5 years.

1.4.1 Disk Storage

Unlike tape storage, disk storage typically has much higher associated costs regarding power and cooling. We also must take into account the reliability and failure rates of disk technology, which limits the typical use of these systems to between 3 to 5 years. To facilitate migration and cost savings we look to double to storage capacity on each migration (e.g., from 1.5 to 3Tb disks)

1.4.2 LTO Tape Storage (Linear Tape-Open)

LTO is a magnetic tape data storage technology, developed in the late 1990s under an open standards initiative as opposed to the proprietary magnetic tape formats that were available at the time. The capacity will approximately double with each new generation (from LTO-1 to LTO-8) but retain the same physical size, thus facilitating migration strategies and allowing standardization in physical storage depots. However, the LTO specification only mandates that each version of LTO is read compatible with the 2 preceding versions (e.g., LTO-4 drives can read LTO-2 tapes) Therefore, to reliably migrate data we must transfer the data to the new format when it becomes widely available and is economically viable, which is approximately every 4 to 5 years.

1.4.3 Checksums

After the creation of the DAFF archive file, its content is verified using various tools. If the content of the file is found to be valid, we take a checksum (SHA) of the file and store this in a separate database.

Checksums are also stored with their associated files when they are stored on magnetic tape. In this way we can periodically verify that the contents of the file are valid and if necessary replace a corrupted copy from a backup.

2. Method

The methodology was as follows:

1. Extract all audio, video and image files (identified by their MIME type) from Ina's archive during the period 1996/7.
2. Use file identification software to verify the content.
3. Assess whether there are any unknown or obsolescent file types.
4. Attempt to read these file types through emulation or migration.
5. Attempt to play these file types by modifying the existing archive interface.

2.1 Small Sample Testing

In order to find the best candidates for emulation/migration we decided to look at the older part of the web archive content collected in 1996/7, which contained approximately 4-5Gb of data. Although this is a very small part of the archive's overall content, we thought it would produce a more manageable dataset. Again, given Ina's interest in audiovisual material we focused on image, audio and video content.

2.2 Hardware used

Desktop: Dell Optiplex 760, 2Gb RAM, Intel Core 2 Duo E8400 3GHz - Windows 7

Data Processing: Dell R710, 16Gb RAM, 2 x Intel Xeon Quad-core X5560 2.8GHz - CentOS 5.5 64-bit + chrooted CentOS 5.5, 32-bit installation

Storage array: Transtec 2 x Intel Xeon Quad-core X5650 2.67GHz, 48Gb RAM, Adaptec 5805Z RAID controller , 2 x 28Tb RAID 6 virtual disks - CentOS 5.5 64-bit

2.3 Format Identifying Software

There are a number of format identifiers available, some well known to the preservation community such as PRONOM and JHOVE. However we decided against using JHOVE in this instance as there is limited support for audio and video file formats.

2.3.1 Format Identification for Digital Objects (FIDO) v1.1.0 (DROID signature file v60)

<https://github.com/openplanets/fido>

Description:

Developed as part of the Open Planets Foundation, FIDO is a Python command-line tool to identify the file formats of digital objects using PRONOM's Digital Record Object Identification's (DROID) signature files.

Comments:

Performed the tests very quickly and was capable of exporting the information to csv format for easy conversion into excel. Backed by the powerful PRONOM library which is capable of using signatures to positively match filetypes (not just by filename extensions). Simple to use and easy to integrate into existing workflows.

Positive Identification Criteria:

We considered a positive match to be where a file format signature was used to identify the object.

2.3.2 Apache Tika v1.1

<http://tika.apache.org/>

Description:

The Apache Tika™ toolkit detects and extracts metadata and structured text content from various documents using existing parser libraries.

Comments:

Tika is also easy to setup and comes with an extensive set of libraries for integration into the user's environment. However for these tests we simply used the included jar file to launch the application.

Positive Identification Criteria:

We considered a positive match where Tika was able to extract significant metadata (i.e., other than the name, size, or content type) from the file.

3. Results

3.1 Extraction Process

The Data processing server ran 4 parallel processes to scan through 270Gb of metadata to extract objects matching the MIME type video/*, audio/* or image/* and the date of capture. From the resulting extraction, we used the content-id checksum in the metadata to link through to the corresponding DAFF data file which could then be downloaded from the archive servers and stored locally on the Data Processing Server.

The extraction took approximately 4 hours and produced the following fileset (Table 2).

Table 2. 1996-7 Archive Sample.

	1996	1997
Files	4607	12206
Size (Gb)	1.4	2.7

3.2 File type assessment

After the extraction we ran both file format identification tools on the extracted files (Table 3). We were surprised at the high success rate of FIDO, which identified the vast majority of formats, whereas Tika struggled with the less common audio and video formats.

Table 3. 1996-7 Positive identification.

Positive identification	1996	1997
FIDO	97.3%	98.3%
Tika	85.1%	87.4%

After using FIDO and Tika we were left to categorise the remaining files (Table 4) without signature based identification (or extensive metadata) by hand. Looking at the list, it became clear that the majority of these files formats are well known, well supported and still in wide use today—with the notable exception of the Vivo Active Video file format.

Table 4. 1996-7 Identified Filetypes.

File Types	1996	1997
GIF 1987a/1989a	1340	3914
Audio Video Interleave (avi)	566	949
JPEG	1447	5169
WAV	580	599
TIFF	0	2
Quicktime (mov)	595	1030
AIFF	31	0
MPEG	0	1
AU	12	3
MIDI	4	15
Real Audio (ra/ram)	32	490
<i>VivoActive (viv)</i>	<i>0</i>	<i>34</i>
Total	4607	12206

3.3 File Format – at risk: Vivo

Out of all the formats identified during this process, only one could be considered as being ‘at risk’—the VivoActive VIVO video format. The format was one of the first used for video streaming in the 1990s before being rendered obsolete by other formats supported by Real Player, Quick Time and Windows Media Player. RealPlayer (Real Networks, n.d.) acquired VivoActive and consequently the VIVO format in 1997.

3.4 Second Archive Sample (1996-2008)

Now that we had identified a particular format of interest, we decided to take a look as to whether other such files existed in the archive. Returning to the metadata we re-scanned files pertaining to a larger subset of the Archive between 1996-2008, this time slightly widening the search to include Vivo’s MIME type (video/vnd) as well as any files matching the filename extension “.viv” (Table 5).

Table 5. Archive 1996 - 2008 Vivo format files.

Year	Vivo files	Year	Vivo files
1996	0	2003	31
1997	209	2004	2
1998	159	2005	1
1999	0	2006	8
2000	17	2007	0
2001	31	2008	0
2002	38	Total	496

We identified 496 files corresponding to these parameters of which 211 were unique (multiple references within the archive). As expected, the usage of this format declined sharply shortly after VivoActive was taken over in 1997.

3.5 Reading the Vivo format

Finding a player for the vivo format proved problematic—there are still links (RealNetworks, Inc., n.d.) to the original software on the Realplayer site, but it appears that any development stopped since the acquisition of the vivo format in 1997—the software is designed for Windows 95 and browsers (Netscape 4 / IE 3 or 4) of that period.

The next port of call was RealPlayer itself (current version 15.0.5.109), however the Vivo format files were no longer supported (Formats - RealNetworks, Inc., n.d.).

In the end, we turned to opensource alternatives such as MPlayer (MPlayer, n.d.) which supports the VIVO format versions 1.0, 2.0, I263 and other H.263(+) variants (using x86 DLLs).

MPlayer is also bundled with Mencoder which allows all readable formats to be converted to modern variants such as MPEG-4 H.264 and Flash Video.

3.6 Problems with supporting Vivo libraries

MPlayer includes support for a very large range of video formats by allowing the inclusion of numerous codecs in the form of libraries. With some of the older formats, including for Vivo, this means using the existing libraries as they were developed, i.e., for windows in the 1990s in the form of 32-bit DLLs.

To use these DLL codecs it is necessary to install or emulate a 32-bit operating system, either through the use of virtual machines or via the use of 32-bit libraries. Given that our server environment is now completely 64-bit linux based, we decided to use a chrooted 32-bit installation of linux on an existing 64-bit machine.

The installation method is too complex to describe in detail in this paper, but essentially this allowed us to preserve all the 32-bit software and libraries necessary to execute Mplayer and Mencoder in this environment.

3.7 Ina's Media Player tool

In the development of the consultation interface for the archive (based on Firefox) we decided to adopt the use of an external media player via a toolbar button (Figure 1) to help play audio and video files in the archive. The media player works by identifying links within the page as well as being able to access metadata which allows it to link media files associated with the page and passes them to an external player (we are currently using the JWPlayer (v5.4) for its enhanced Flash and HTML5 capabilities).

In one example, embedded flash videos on a page are captured separately by data mining the links on the archived page and downloading the video content in a separate procedure. In Figure 2 we can see how the archived web page is missing its embedded video. In Figure 3 we can see how the video content has been 're-associated' with the page by using the media player plugin.

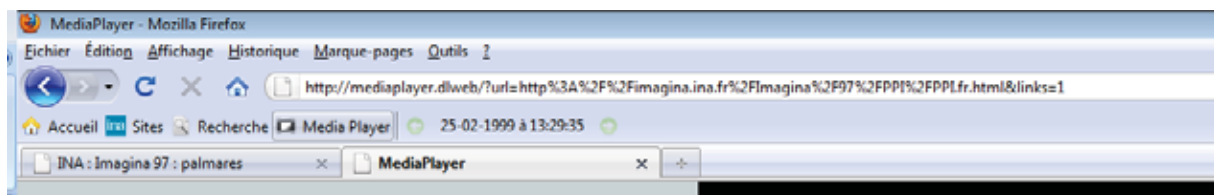


Figure 1. Firefox Media Player Toolbar Button.

3.8 Integrating Mplayer/Mencoder with Media Player

Using the media player plugin and combining it with Mencoder we are able to convert 'on the fly' associated vivo files to a more compatible format. In this case we have chosen the Flash file (flv) format.

The vivo video format is identified through its MIME type (video/vnd.vivo) and converted automatically by Mencoder. The converted files are then stored in their new format and can be played directly (without the need for conversion) the next time the video is requested. In essence, we are performing 'on-demand' migration.

In Figure 4 we have a webpage with links to vivo files. Using the modified Media Player we can access the automatically converted content (Figure 5).

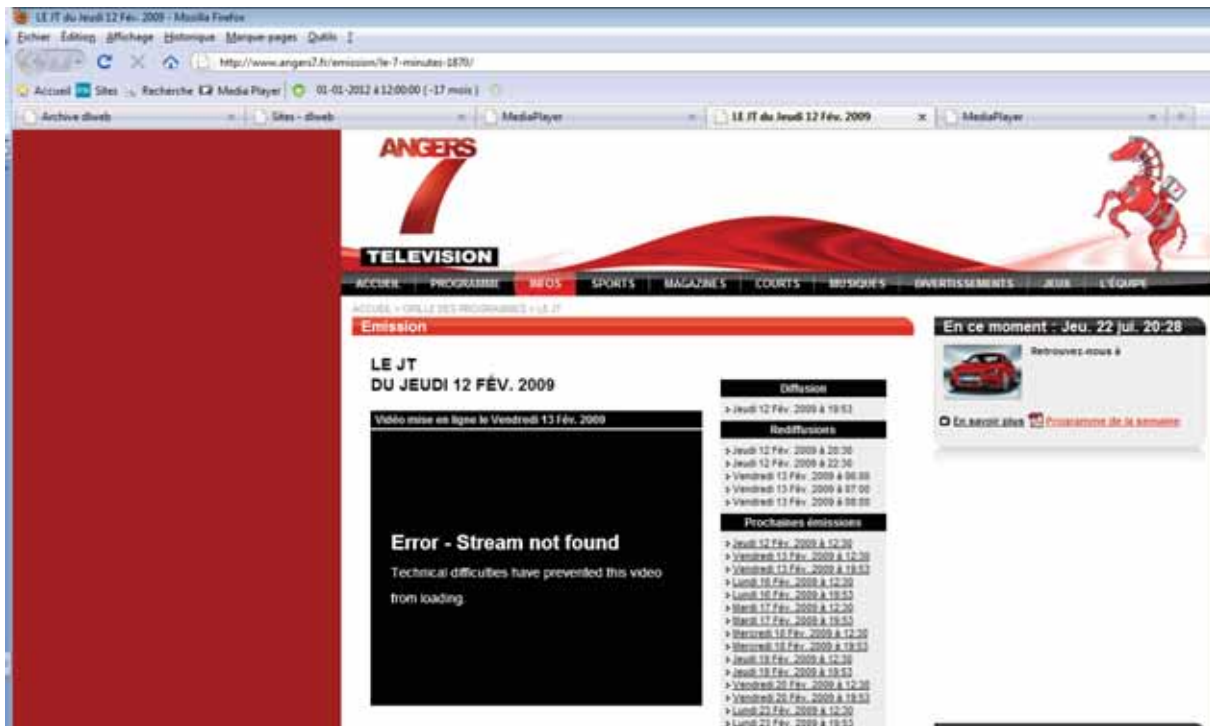


Figure 2. Archived Angers 7 page missing embedded video content.

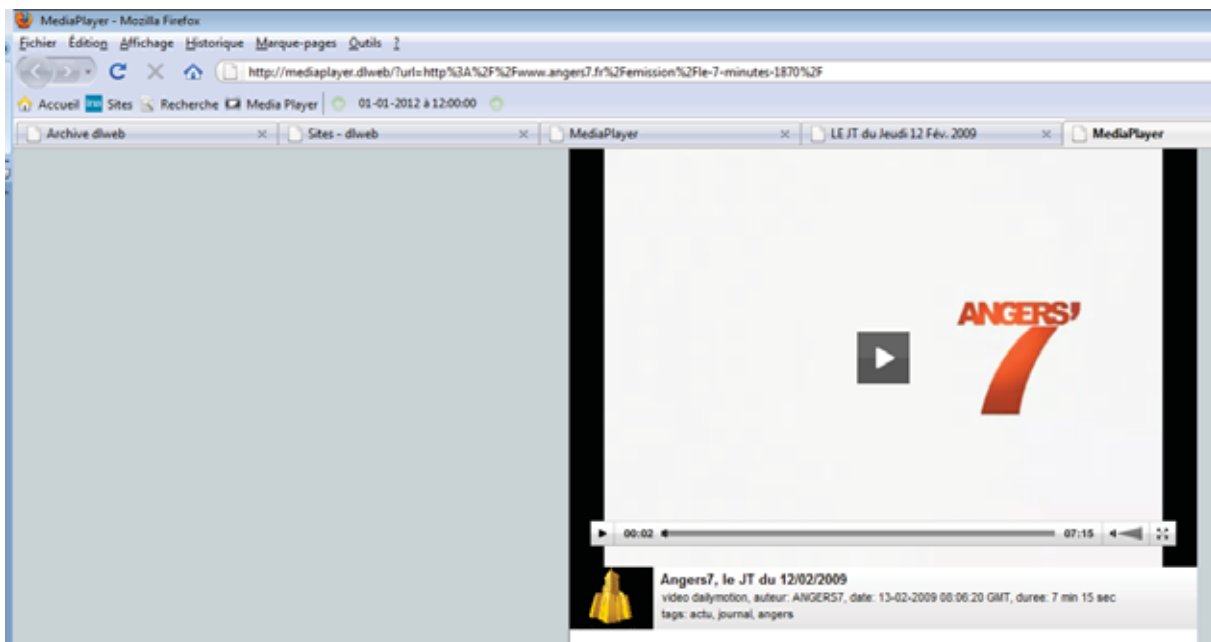


Figure 3. Archived Angers 7 page recombined video content accessible through Media Player.



Figure 4. Web page with vivo video links.

3.9 Conversion Compatibility Problems

It should be noted that we had some difficulty in choosing options with which to convert the files. There are a wide range of options available in terms of bit rates, size, quality synchronization, etc., which changed the length and quality of the resulting video considerably. Further investigation is required on how to better normalize the process.

4. Conclusions

In the end we were able to use transparent format migration, underpinned by some emulation, to help us access an obsolete file format. This work was very much a test of the technology, able to be built on to the existing structure of the archive (Media Player) and we were able to see how well the process could be integrated. This has resulted in a practical tool which now enables users of the archive to access videos which had been effectively unreadable.

That said, the ‘obsolescence’ of the Vivo file format could be considered relatively insignificant given the very small percentage of files as compared to the size of the archive. However, given that a large percentage of these files were found on old archived pages from Ina’s corporate website, they take on a much greater importance!

4.1 Further Work

This work opens up some interesting avenues to pursue with future projects. With the increasing use of data mining on our own archive, we can imagine conducting file format identification on the whole

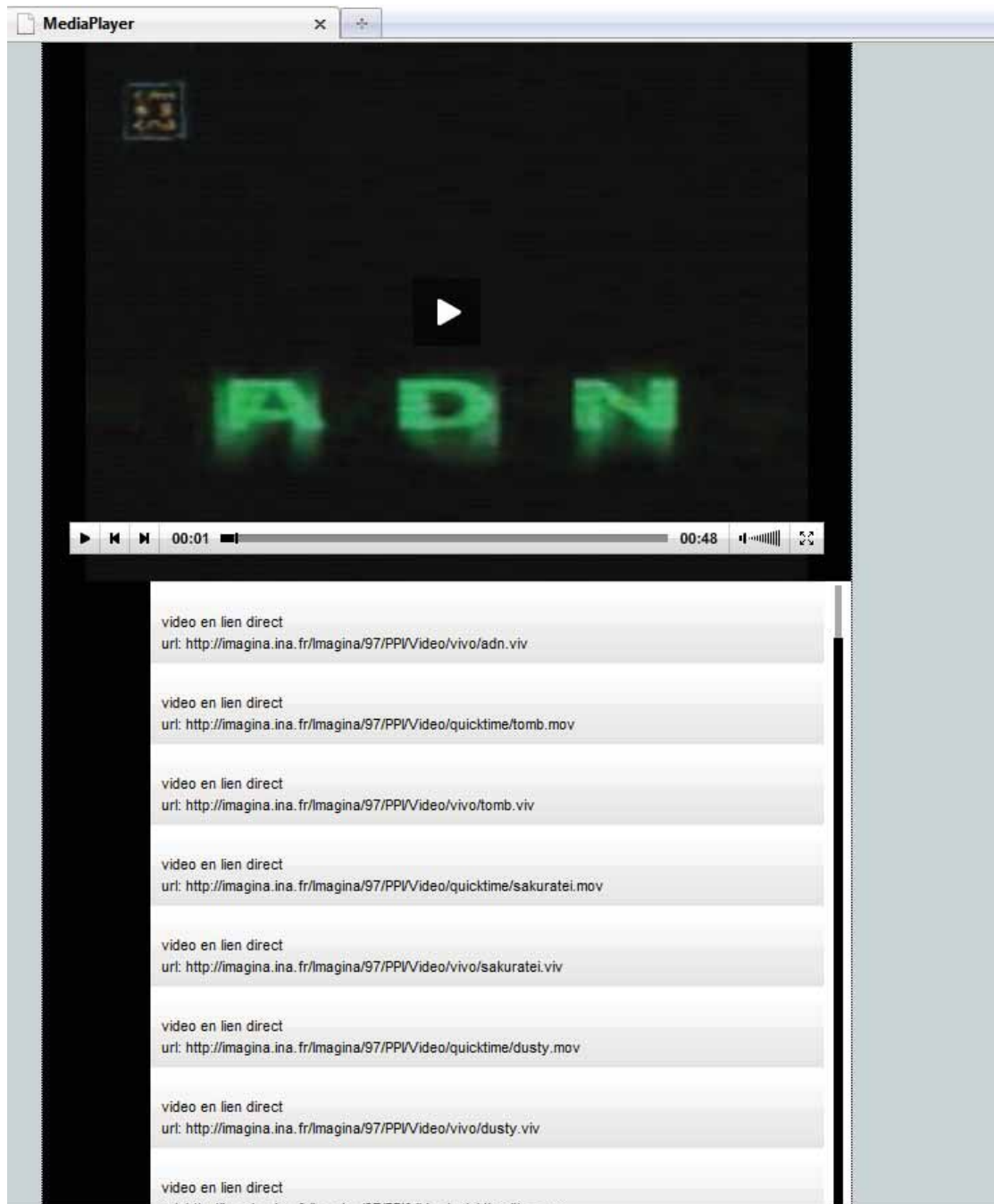


Figure 5. MediaPlayer playing videos converted automatically from Vivo to Flash video format.

archive to begin to analyse trends in file format obsolescence. The tools seemed to be robust and useable on a much wider scale.

Even though MPlayer/Mencoder is a modern up to date piece of software, we are still relying on emulation in the form of codecs and a 32-bit environment upon which to play the videos. We need to look

at ways of cataloguing and preserving such information so it can be used to help the emulation of software or to support renderers in the future.

All of the work presented in this paper will be fed back into the Preservation Working Group of the IIPC, where a number of projects are underway to help the web archiving community track format obsolescence as well as the tools to support them.

4.2 Looking Forward

Preserving a web page in the archive, as it was, will become increasing complex task as software develops to keep up with the myriad of formats and standards. Although we believe we will be able to preserve a lot of the text, images and videos we think there will inevitably be a loss with regards the style, feel and interactivity of web pages over time. Already with the use of the Media Player we are, in essence, breaking up the flow of a web page to be able to extract the content.

We think it is best that we should concentrate on using current tools to extensively data mine as to enrich the archive with as much metadata as possible. In this way we should look to leave as many tools as possible for future generations to re-interpret these digital artefacts.

References

- Brown, R. H., and B. Davis-Brown. "The making of memory: the politics of archives, libraries and museums in the construction of national consciousness." *History of Human Sciences* 11, no. 4 (1998): 17-32.
- Carr, E. H. *What Is History?* New York: Random House, 1961.
- Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). 650.0-M-2. 2012.
- Cox, R. J. "The Documentation Strategy and Archival Appraisal Principles: A Different Perspective." *Archivaria* 38 (1994): 11-36.
- Diamond, E. "The Archivist as Forensic Scientist—Seeing Ourselves in a Different Way." *Archivaria* 38 (1994): 139-154.
- Duranti, L., and B. Endicott-Popovsky. "Digital Records Forensics: A New Science and Academic Program for Forensic Readiness." *Journal of Digital Forensics, Security and Law* 5, no. 2 (2010): 45-62.
- Formats - RealNetworks, Inc . "Supported Media Formats." (n.d.). Retrieved from http://cache-download.real.com/free/windows/mrkt/help/RealPlayer-15/en/Content/Media_Types.htm.
- Gardner, J. "The Mark of Responsibility." *Oxford Journal of Legal Studies* 23, no. 2 (2003): 157-171.
- Halbwachs, M. *On Collective Memory* (originally published in 1941, edited and translated by Lewis A. Coser). Chicago: University of Chicago Press, 1992.
- Hedstrom, M. "Exploring the Concept of Temporal Interoperability as a Framework for Digital Preservation." In *Third DELOS Network of Excellence Workshop on Interoperability and Mediation in Heterogeneous Digital Libraries*, 2001.
- Josias, A. "Toward an understanding of archives as a feature of collective memory." *Archival Science* 11, no. 1-2 (2011): 95-112.

- Lee, C. A., and H. R. Tibbo. "Where's the Archivist in Digital Curation? Exploring the Possibilities through a Matrix of Knowledge and Skills." *Archivaria* 72 (2011): 123-168.
- Liu, W. W. "Identifying and Addressing Rogue Servers in Countering Internet Email Misuse." In *IEEE/SADFE-2010: Proceedings of the 5th International Workshop on Systematic Approaches to Digital Forensic Engineering*, 13-24. Oakland, CA: IEEE Computer Society, 2010.
- Liu, W. W. "Trust Management and Accountability for Internet Security." Ph.D. diss., Florida State University, 2011.
- Liu, W., S. Aggarwal, S., and Z. Duan. "Incorporating Accountability into Internet Email." In *SAC'09: Proceedings of the 24th ACM Symposium on Applied Computing*. Honolulu, HI: ACM, 2009.
- Long, A. S. "Long-Term Preservation of Web Archives – Experimenting with Emulation and Migration Methodologies." 10 December 2009. Accessed 1 April 2012, http://netpreserve.org/publications/NLA_2009_IPC_Report.pdf.
- Misztal, B. *Theories of Social Remembering*. Maidenhead, UK: Open University Press, 2003.
- MPlayer. (n.d.). Retrieved from <http://www.mplayerhq.hu/>.
- Nora, P. "Between Memory and History: Les Lieux de Memoire." *Representations* 26, sp. issue (1989): 7-24.
- Pruzan, P. "From Control to Values-Based Management and Accountability." *Journal of Business Ethics* 17, no. 13 (1998): 1379-1394.
- Real Networks. "RealNetworks, Inc. - Acquisition History." (n.d.). Accessed 7 May 2012, <http://investor.realnetworks.com/faq.cfm?faqid=2>
- RealNetworks, Inc. "Download the VivoActive Player." (n.d.). Retrieved from <http://egg.real.com/vivo-player/vivodl.html/>.
- Rosenthal, D. S., T. Lipkis, T. S. Robertson, and S. Morabito. "Transparent Format Migration of Preserved Web Content." 1 Jan 2005. Accessed 7 May 2012, <http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html>.
- Sachs, A. "Archives, truth and reconciliation." *Archivaria* 62 (2006): 1-14.
- Schwartz, J. M., and T. Cook. "Archives, records, and power: the making of modern memory." *Archival Science* 2, no. 1 (2002): 1-19.
- Sinclair, A. "The Chameleon of Accountability: Forms and Discourses." *Accounting, Organizations and Society* 20, no. 2/3 (1995): 219-237.
- Wallace, D. A. "Introduction: memory ethics—or the presence of the past in the present." *Archival Science* 11 (2011): 1-12.
- Wallace, D. A., and L. Stuchell. "Understanding the 9/11 Commission Archive: Control, Access, and the Politics of Manipulation." *Archival Science* 11, no. 1-2 (2011): 125-169.
- Yakel, E. "Digital Curation." *OCLC Systems & Services* 23, no. 4 (2007): 335-340.
- Yakel, E., P. Conway, M. Hedstrom, and D. Wallace. "Digital Curation for Digital Natives." *Journal of Education for Library and Information Science* 55, no. 1 (2011): 23-31.

Technology as the Mediator of Heritage and Its Relations with People

The Turtle At The Bottom

Reflections on Access and Preservation for Information Artefacts

Ian S. King

University of Washington

Abstract

Issues of access and preservation for digital documents transcend the use of digitization to refactor originally non-digital content. Born-digital documents are contextual within their originating digital environment. Preservation of the originating execution environment maximizes preservation of that context. The two common strategies for preserving the execution environment—maintaining running systems, and software emulation of such systems—are interdependent mechanisms. Thoughtfully decomposition of original information hardware systems and isolated emulation of risk-prone components preserves crucial primary-source references to validate authenticity of software emulations, upon which both born-digital and transcribed-to-bits documents may be faithfully preserved and widely accessed.

Author

Ian S. King is a second-year doctoral student in The Information School (iSchool) at the University of Washington in Seattle. He previously earned a Master's of Science in Computer Science from UW in 2006. King works as a senior systems engineer and curator at the Living Computer Museum, a presentation of Microsoft co-founder Paul Allen's collection of vintage computer systems. He has worked with computer based information systems for many years, having learned FORTRAN as his first programming language in 1974, coding on punched cards. His working life has also included technologies such as radio communication and other electronic systems.

1. Introduction

There is an old joke in which an elderly person refutes a scientist's account of cosmology, stating firmly, "The Earth is balanced on the back of a great turtle." The scientist asks, "Respectfully, what is that turtle standing on?" Querulously, the elder replies, "Of course, it's turtles all the way down!"¹ In many discussions of preservation of digital documents, writ large, this author observes the (hopefully unintended) construction of similarly perilous dependency chains, some unbounded and some simply fragile. This paper will frame a discussion of the problem and discuss a dialectic of current thinking for solutions, with particular attention to new challenges posed as the digital document evolves. Further, I will present a case study of a strategy that addresses some of the most vexing issues in both classic approaches.

This discussion begins with the classification of the subject matter, that is, digital documents. The emergence of conversational computing has engendered a subset of so-called "born digital" documents that arguably cannot be presented without the context of their creation. I suggest a realignment of some commonly used terms, *static* and *dynamic* documents, to better frame an ontology of conversational computing artefacts.

¹ I mean no disrespect to anyone's views of cosmology with this Western-Enlightenment-premised joke.

If these dynamic digital documents require the context of their creation, how shall scholars address this dependency in a manner that can support access to and preservation of these documents over a course of decades or centuries? The two approaches most commonly discussed are the maintenance² of the original systems in working order, and the emulation of those systems in software that executes on today's hardware and software platforms. This paper offers an overview of the structural natures of these approaches, together with strengths and weaknesses. This is followed by an overview of a third approach that incorporates ideas from both strategies. This third path arguably offers greater assurances of integrity and reliability over the long-term for both dynamic and static digital documents.

This third path is presented as a case study performed in a somewhat ethnographic fashion, based on the author's work since 2008 with the Living Computer Museum in Seattle, Washington (LCM). I disclose and make no excuse for my dedication to the maintenance of vintage computer systems in running condition, and in this paper seek to demonstrate that such efforts are critical during a transitional period of uncertain but discernable duration.

2. Framing the Problem

2.1 Understanding the Turtle at the Top

Patricia Battin has been cited as stating, "Access is preservation, and preservation is access."³ As noted by other authors,⁴ this is often taken out of context and interpreted as a statement of equivalence. Battin was discussing the benefits of digital transcription and preservation for "crumbling books" (in the original) and can be more comprehensively interpreted: this author reads the broader text as a statement that access without preservation is catastrophic for the book, and preservation without access is a hollow act. However, are these concepts as currently framed sufficient to discuss digital documents in their fullest sense?

Our use of digital technology for the creation, retention and dissemination of documents falls into two large areas. As Battin describes, we are transcribing documents created with earlier technologies into collections of bits. A crumbling book can thus be shared with a broader audience without fear of damaging the original. The content is made accessible, while the original artefact is preserved. We are digitizing books, images, sound and video to this end.

The second broad scope is commonly called "born digital" documents. These artefacts have never known physical form, but are created and persist as collections of bits. Blanchette argues the point that there is in fact a manifestation of these bits, or they cannot be either preserved or presented.⁵ He also states that, as digital bits, they are reproducible without error and therefore the constructs of bits are effectively eternal. There is a dilemma here, however: if a collection of bits resides in some physical form that subsequently degrades, those bits may be lost. The eternal life of the bit promised by Blanchette can perhaps be better characterized as an opportunity for reincarnation, whether the bits are regenerated in their existing manifestation or migrated to another carrier.

² I consciously employ the term 'maintenance' rather than the appealing term 'preservation,' which is often interpreted differently by archivists than I would use it here.

³ Patricia Battin, "From Preservation to Access: Paradigm for the Nineties," *IFLA Journal* 19, no. 4 (1993): 367-373.

⁴ E.g., M. V. Cloonan, "W(h)ither Preservation?" *The Library Quarterly* 71, no. 2 (2001): 231-242.

⁵ Jean-François Blanchette, "A Material History of Bits," *Journal of the American Society for Information and Technology* 62, no. 6 (2011): 1042-1057.

At this level, this sounds like simple media migration. But the issue is far more involved and perhaps intractable for some forms of the digital document.

Orthogonal to the above analysis, Rothenberg distinguishes two types of digital document, *static* and *dynamic* documents, by categorizing text and still images as the former, and multimedia (such as a video presentation and web pages) as the latter.⁶ When considering a broader view of the digital document, these definitions are limiting: I argue that the distinction is better drawn as follows. If I read a book, I expect the words to be the same each time I turn to a particular page. This is true even if the book is preserved as a string of bits: we measure the quality of a rendering mechanism by the fidelity to the original text. This is also true for still images, audio streams and video. On subsequent renderings of “Downhill From Here” by the Grateful Dead, I expect that Jerry will play the same notes and sing the same words, and that “Space Drums” will go on forever. Each presentation must be identical to previous presentations, or we declare the rendering to be false. These are all static documents.

Contrariwise, a computer video game interacts with its human participant to create a unique experience with each interaction. Other examples of this idea are computer-based data visualization and computer-supported collaborative work (CSCW), and social networking can be viewed similarly. One can write papers about a particular game, provide still photos or even video clips demonstrating game play, but these records do not provide a faithful reflection of the conversational experience with a game system: you’ll never know what it feels like to turn your spaceship and fire, madly mashing buttons, only to be destroyed by the asteroid coming at you from your blind side (i.e., the game “Asteroids”). These are better described as dynamic documents.

In other words, there is a class of digital documents that include elements that must be experienced interactively to convey meaning. This class demonstrates the convergence of computer technology with human agency in what J. C. R. Licklider foresaw as a sort of symbiosis.⁷ Early information systems produced an end result to a given interaction, and meaning was ascribed to that result: if there is a certain symbol at the beginning of a printed string of digits, this means the balance of your checking account is negative and you are in trouble. Further along this evolutionary arc, meaning moved to the interaction itself: there is no similar end result to a video game or a computer-mediated social interaction. I refer to such information systems as *conversational*, as ‘interactive’ has been interpreted much more broadly (and probably irrevocably) as the diametric concept to batch processing.

Ensuring preservation of and access to dynamic digital documents is the focus of this discussion. The benefits that may inure to the preservation of static documents, including those transcribed from analogue form, is a fortuitous side effect.

2.2 Another Stack of Turtles: Bits Are Not Fungible

There is a confidence to preservation of digital documents that is premised on the nature of digital content, that it is composed of bits. Bits can be regenerated without error, and are effectively eternal, goes the argument.⁸ However, bits are just that: bits. Collections of bits have no meaning except that which human agency provides. Beyond the bits themselves, we therefore must create or preserve metadata to

⁶ J. Rothenberg, “Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation,” 1999.

⁷ J. C. R. Licklider, “Man-Computer Symbiosis,” *IRE Transactions on Human Factors in Electronics* 1, no. 1 (1960): 4-11.

⁸ Blanchette, *supra*.

allow for the reconstruction of meaning. The metadata most often exists at several levels, all of which must be preserved or meaning is lost.

For example, Digital Equipment Corporation's DECTape magnetic tape storage system⁹ records patterns of magnetism in ten parallel tracks. Those patterns are encoded in a fashion that may be recovered as bits. Two tracks record bits representing timing information, making the system robust against variations in transport speed. Two others tracks record bits representing metadata that establishes data blocks on the tape and enables an encoding to define the meaning of a given block within a group of blocks. The other six tracks encode content as bits in a redundant fashion: each bit of groups of three is represented twice in the same lateral location. Above this level, a file system may be imposed upon the blocks, and the blocks may be of arbitrary length (within limits of word size). With all of this being established, a program such as an operating system may now define files that mean... whatever we said they meant when they were created.

There has been substantial discussion by others of the selection and/or generation of appropriate metadata to accompany various types of digital documents. And how shall we represent and preserve that metadata? It says right here that "bits can be regenerated without error"—why don't we do that? Great idea, but there's one problem: those bits have no inherent meaning, either. So we need metadata to describe our metadata, and... the turtles just keep stacking up. There is no implicit bound.

2.3 When Turtles Won't Talk To You: Proprietary Formats

Another obstacle is the use of proprietary document formats, which may preclude the production of meaningful metadata to enable rendering.¹⁰ As open source document formats proliferate, it seems possible the issue of proprietary formats may be less of an issue in the future, but this does not address the past in which such mechanism were common. The two apparent solutions are migration to open formats and preservation of execution environments. The former risks data loss on transcription and, even if great care is taken to preserve the apparent rendering, there is no guarantee that unrendered information is not lost. For example, some word processing formats preserved a number of edits in the data file, enabling the ability to observe the evolution of a document.

2.4 The Turtles (May) Stop Here

Let us assume media on which is stored a set of bits representing a digital document. The goal is to present the document authentically and meaningfully, in a way that one can consume it with the senses and interpret it within one's reality.

An obvious method to ensure authenticity is to associate the media with the platform of its origin to render the document. Logically, employing the historically correct platform for the presentation of a digital document ensures an authentic experience insofar as the presentation elements of the experience (as distinct from interpretation within the observer's reality, which is not a technology issue).

Another method, one that has gained popularity in recent years, is software emulation of the original system. Emulated hardware enables the original software environment to be employed. If the bits of the

⁹ Leonard M. Hantman, "Microtape: Its Features and Applications," in *Second Annual Meeting, DECUS* (Maynard, MA: Digital Equipment Corporation, 1963). Note that Microtape was the original name for DECTape.

¹⁰ J. R. van der Hoeven, "Development of a Universal Virtual Computer (UVC) for Long-term Preservation of Digital Objects," *Journal of Information Science* 31, no. 3 (2005): 196–208.

document and its supporting software can be transcribed from original media with fidelity, a well-designed emulation can provide a substantially faithful presentation of the elements of the digital document.

In both cases, we may have found the turtle at the bottom: the operating environment in which the digital document was born. Such a mechanism—whether physical or emulated—subsumes the metadata requirements for its layers of meaning. If our document is rendered on either a vintage computer or its emulation, *assuming* adequate fidelity of the emulation (and complete preservation of its bits), our task of understanding begins with the document itself.

There are still substrata for either of these turtles, and a key reason for this discourse is to analyse just how ultimate our bottom turtle may be considered. The following sections will discuss each approach, with strengths and weaknesses noted.

2.5 Maintaining Running Systems

The preservation of vintage computer hardware together with its software challenges the language we use to describe it. A classic archival museum employs the term ‘preservation’ to describe the maintenance of an artefact in its current condition over an indefinite span of time. When such a museum undertakes a ‘restoration’, the goal is to address previous degradation or damage, again with the goal of presenting the artefact in that restored condition.

The Living Computer Museum stands with a significant number of private collectors of vintage computer systems in its goal to preserve information systems as functional devices. This often involves repair of a system that is no longer functional and maintenance thereafter to maintain functionality. There is disagreement in the computer museum community as to how history is best served, but this paper is premised on maintaining functional systems for as long as possible.

2.6 Structure of Running Systems

Electronic digital computer systems were first constructed using thermionic vacuum tubes (called valves in some locales), which were later replaced by transistors. The technology of the discrete transistor soon evolved to allow multiple transistors on a single substrate, and they were combined with other components to create complete “integrated circuits”. The transistor density of integrated circuits has grown dramatically with time, allowing major portions of a computer system to be collocated on a single physical component.

Regardless of physical manifestation, the elements of such systems map onto the radial model described above. A vintage mainframe such as an IBM 7094 is expressed in several large cabinets, while a laptop computer may be so highly integrated that its components cannot feasibly be removed for replacement. But for the majority of the history of computer technology to date, the elements upon each arc of functionality can be physically distinguished.

Some types of media used with running systems must be considered consumables in operations planning. Magnetic disks and flash drives can be purged and used again, but media such as punched cards and paper tape are “write once” media. Unlike visual displays, Teletypes and printers consume paper for each row of output.

2.7 Strengths and Liabilities

Regarding preservation: the presentation of a digital document in its original environment is beyond reproach in its authenticity. This environment may include both essential hardware and specialized elements, for example stereoscopic visual displays or physical human interaction devices such as joysticks.

Regarding access: maintenance of running systems presents challenges of structural accessibility, which will be described in the following paragraphs.

Early systems consumed a great deal of electrical energy (in some cases on the order of kilowatts) and did not make efficient use of it, requiring yet more energy expenditure for cooling. Physical space is another cost. Mainframes required considerable space both for operation and maintenance access. Although so-called minicomputers were significantly smaller than their predecessors, there was also a great diversity of models: a working environment might rely on one model, but the needs of a digital repository could easily require numerous systems.

The enabling technologies often involved electromechanical elements that inevitably wear with time. Mass storage devices are traditional challenges, but other items such as cooling fans and switches also fail mechanically. The issues of media have been discussed at length elsewhere.

Another challenge not anticipated by the originators of these systems is failure of purely electronic components. One especially weak link is the aluminum electrolytic capacitor, which is discussed below (Short Stories: Power Systems). Research suggests that some classes of integrated circuits fail due to migration at a molecular level.¹¹ Failure of internal storage have been observed to affect field-programmable active components such as gate arrays; most often, these devices' programming cannot be read from the device and the original programming sources have been lost.

To date, there is considerable success in restoring systems from the 1960s through the early 1980s. The earlier systems can be simpler to restore because of their use of discrete transistors, for which modern substitutes are readily available. Into the 1970s, the widespread use of integrated circuits is problematic because such devices are no longer manufactured. In the 1980s, the rise of application-specific integrated circuits (ASICs), while reducing production cost, renders devices irreparable except through the scavenging of components from multiple systems to construct one working model. The rapid evolution of complex integrated circuits is evidenced by short product lifecycles, suggesting that the challenge of acquiring replacement parts will continue as today's systems become tomorrow's history.

Regardless of the successes in maintenance of running systems, another simple limitation is their existence. Some systems exist in single-digit quantities.

3. Software Emulation

3.1 Structure of Software Emulations

A software emulation is a computer program that is executed on a *host platform* that is most often dissimilar from the *emulation target*. It consumes the instruction stream of original software to replicate its transformations of a state of computation, represented in data structures rather than hardware registers. The use of software emulation (or sometimes, simulation) has a long history: it was common practice to

¹¹ P. Yalamanchili and A. Christou, "Migrated copper resistive shorts in plastic encapsulated devices," *Integrated Reliability Workshop, 1995. Final Report., International*, p. 166.

design new computers as software emulations on existing machines. Over the past several years, software emulation has become an oft-discussed tool for presentation of digital documents.

Note that binary translation methods are not discussed here.¹² Such techniques transform the instruction stream, while the focus of this paper is on strategies that address the entire execution ecosystem of a digital document.

Commonly, many peripherals of the emulation target are mapped onto those of the host platform: for example, if the host is a PC, its keyboard and display often serve the function of a video display terminal. Character-cell peripherals are often emulated with console applications. Those emulation targets that include graphical display employed either a vector display or a bit-mapped, all-points-addressable display. The former is solved with a well-understood vector-to-raster conversion,¹³ and the latter requires simple orthogonal translation. Depending on how the emulation is structured, it may communicate directly with mass storage on the host platform, or (more commonly) mass storage may be modeled in the host's working store and persisted on the host 'out of band'. Working store often does not map directly onto the host, either: for example, the PDP-8 is a 12-bit architecture while the PDP-10 employs a 36-bit word.

The majority of software emulations present instances of all system components as a *monolithic* whole. The structure of the emulation may be *functional*, in which elements of the target are modeled as "black boxes" at various levels, or *architectural*, with models of circuit elements assembled in a reductionist strategy. As a matter of coding practice the elements are almost always separated in distinct code modules, but the code is compiled to a single executable. Amending the emulation, e.g., adding a new peripheral, requires programming skill.

When modeling systems with few configuration options such as a game console, the emulation designer often hard-wires the set of choices. An emulation of a general-purpose target such as a minicomputer usually allows for configuration choices, electing which system components are available to software. The interaction with the host system varies: some emulations include features allowing files to be moved onto emulated mass storage from the host's file system or interaction with physical terminal devices.¹⁴ Others allow little or no exposure of any emulated device outside the emulation.

Software emulations are typically frame-based, modeling the original hardware by transitioning state with each processor cycle. A serious challenge of this strategy is the inherently concurrent nature of non-processor elements. For example, in a physical system a disk drive may be seeking to a cylinder over a number of milliseconds, enough time for hundreds of processor cycles that are concurrently changing the system's state. Networking is particularly problematic, as incoming data is asynchronous to the receiver at many levels.

Software emulations are created in the programming language of the designer's choice. Given the need to model not only the transition of processor state but also potentially concurrent peripherals and storage devices, designers choose languages that compile to time-efficient code: the C programming language is a common choice. Assembly language is rarely used, to avoid tying an emulation to one host platform. Single-threaded emulations are the rule: although a multithreaded solution seems ideal for the

¹² Cristina Cifuentes and Vishv M. Malhotra, "Binary Translation: Static, Dynamic, Retargetable?" in *Proceedings of the 1996 International Conference on Software Maintenance* (Washington, DC, USA: IEEE Computer Society, 1996), 340–349.

¹³ W. M. Newman, "Trends in Graphic Display Design," *IEEE Trans. Comput.* 25, no. 12 (December 1976): 1321–1325.

¹⁴ John Wilson, "Ersatz-11," 1998, <http://www.dbit.com/>.

concurrency described above, general-purpose operating systems are ill-suited to ensuring timing requirements among threads of execution. The cure is worse than the disease.

Some work proposes to create a universal language to express the elements of an information system, so that emulations may be written once and only the underlying platform, i.e., the universal language, needs to be ported to new hardware and/or operating systems.¹⁵ However, once again we are adding at least one more turtle: while the expression of these primitives are argued to be more straightforward, this requires additional proof and evaluation.

3.2 Example: SIMH

The SIMH emulation platform is well known and widely used by enthusiasts as well as those supporting legacy software.¹⁶ There are over two dozen emulations of vintage computer systems available on the SIMH website.¹⁷

SIMH is structured as a tool kit written in the C programming language, available on most host platforms that support C. It offers infrastructure to support a frame-based emulation. Numerous emulated systems are available in both source code and compiled versions for common host platforms.

In order to make use of a particular emulation, it is necessary to configure its functional elements from a set of choices presented as a command-line syntax. One chooses features in the central processor, the quantity of working store, size and kind of mass storage and peripheral devices. The command-line console window in which the emulation is started serves as a terminal device, that is, primary user interface.

Tools have been created that can ease the task of representing files on the host platform so they can be accessed by the emulation from the host. Ultimately, however, a ‘bridge’ must somehow exist between a physical vintage system and a modern system in order to migrate the bits into an accessible form. Such migration work will almost always require preservation in running condition of at least one example of the target system.¹⁸

3.3 Strengths and Liabilities

Regarding access: well-designed emulations, such as the SIMH collection, can provide a substantially authentic experience of original systems. (For the SIMH examples, this is facilitated by the fact that most of the emulated systems presented primarily through character-based interfaces.) Tools allow simple emulation of mass storage devices. The only criterion for a host platform is that is sufficiently performant: emulation typically requires multiple host-level instructions to effect the execution of a single target level instruction.

Regarding preservation: the history of such emulations demonstrates the challenge of correctly preserving the execution environment and, thus, the authenticity of the dependent digital document.

¹⁵ Raymond A. Lorie, “Long-term Preservation of Digital Information,” in *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries* (New York, NY, USA: ACM, 2001), 346–352.

¹⁶ It is published under an open source policy. Many other emulations are licensed for free use by hobbyists or academics, but bear licensing fees for commercial uses.

¹⁷ Robert M. Supnik, “The Computer History Emulation Project,” <http://simh.trailing-edge.com/>.

¹⁸ One LCM engineer did effectively navigate the entire ‘stack’ of meaning from images of magnetic domains on a tape, through several layers of formatting and to the recovery of recognizable data files. It was not an easy task and was supported by extensive documentation and prior knowledge of the expected content.

As an example of the liability of emulation, what follows is an analysis of the mature and respected SIMH platform, through the change log embedded in each release. A textual analysis of the current change log, encompassing a period from October 2001 through May 2012, demonstrates 1,100 lines of changes not attributable to infrastructure (determined by excluding the ‘scp’ sources). The word ‘fixed’ appears in 449 lines and the phrase ‘fixed bug’ in 181 lines; random sampling of other lines shows many being new or extended features, but some representing ‘fixed bugs’ without use of those words. An active Internet mailing list supporting SIMH demonstrates ongoing discovery of issues related to both implementation of target system features and failures due to host platform changes, many discovered through comparison with original systems. SIMH is a well-written platform that has been actively used for over ten years.

There is no intrinsic connection between the execution environment of an emulation and the host platform. This enhances portability and therefore access, but allows an arbitrary relationship between the emulated platform and the underlying host. This relationship is mediated by the host’s instruction set architecture, language processors (compilers) that generate emulation executables, system libraries that support those executables, and the host operating system, including how its device drivers virtualize hardware.

Changes in these levels cause turtles to lose their footing in the tower. The code written by the emulation designer may be compromised in its meaning through compiler changes, library changes or changes in the mapping between software and hardware data types, e.g., 32-bit vs. 64-bit processors.

It is tempting in writing an emulation to employ a functional approach to combine elements that are rarely exposed individually. The author discovered the down side of this approach with the SIMH VAX-11 emulation, attempting to validate diagnostic disk images. The boot disk interface present on a physical VAX-11/780 was not implemented on the SIMH emulation.

A paper addressing the testing of software emulations found “several defects” in four emulations of the modern IA32 architecture, “some of which can prevent the proper execution of programs.” Testing was dependent on comparison of the behavior of an emulation with the physical device being emulated. The authors emphasize the challenge of writing a CPU emulation, and their work highlights the pitfalls in detail.¹⁹

3.4 Summary: Facility for Access and Preservation

Software emulations can be designed to offer broadly available access to the original execution environment and, therefore, a broad range of digital documents. Access is compromised when the presentation elements of the target system cannot or are not interpreted and implemented in a manner that ensures authentic rendering of the elements of the document.

Preservation of the execution environment and, therefore, the digital document as a dependent of that environment, is difficult in the first instance and fragile over time. Even a well-designed emulation faces challenges from changes in platform elements. Efforts²⁰ to transcend these issues are welcome but seemingly add more turtles to the tower, as future digital librarians must reproduce the work of the meta-platform creators.

¹⁹ Lorenzo Martignoni, Roberto Paleari, Giampaolo Fresi Roglia, and Danilo Bruschi, “Testing CPU Emulators,” in *Proceedings of the Eighteenth International Symposium on Software Testing and Analysis* (New York, NY, USA: ACM, 2009), 261–272.

²⁰ Lorie, *supra*.

4. A Third Path: Decompose and Emulate

Software emulations are accessible but there are legitimate concerns about their long-term integrity. Maintaining original systems in running condition precludes questions about the fidelity of documents and preserves them reliably but there are serious problems with accessibility: some systems exist only in single-digit numbers and their maintenance is not for the meek (or poorly-funded).

As the Living Computer Museum (LCM) has pursued a path to maintaining running systems, with a strong commitment to preservation but a mandate committing the organization to access, pragmatic choices have led to exploration of emulation of the most risk-prone *components* of an otherwise original system. To identify these components, LCM engages in thoughtful decomposition of the information hardware.

4.1 Decomposing the information system

The following describes a structural model for the hardware of an information system to facilitate this discussion. Note that this discussion does not include the software component of an information system.

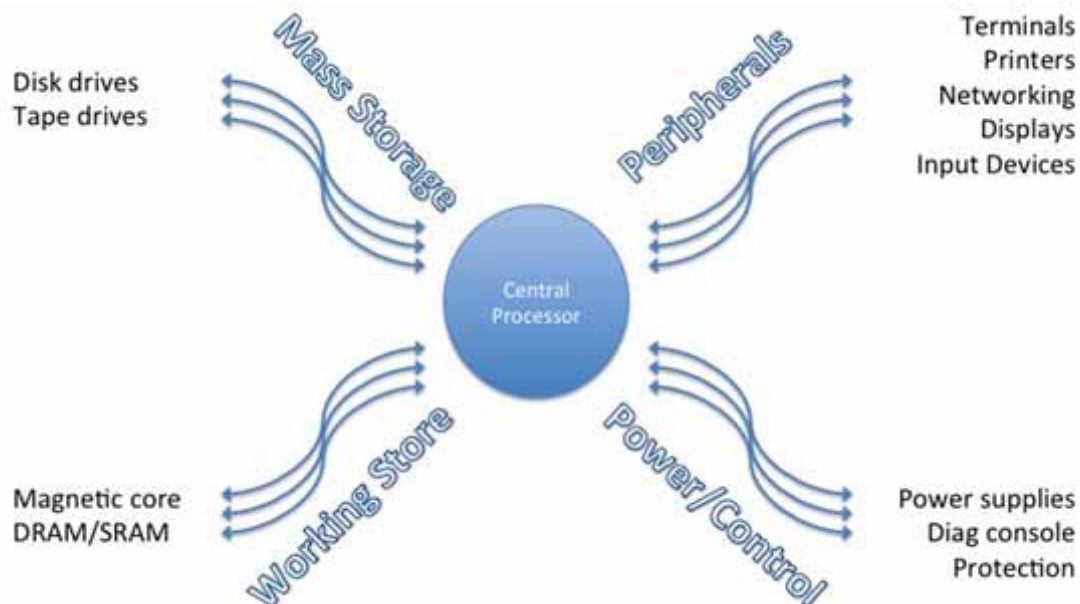


Figure 1. Radial model of a computer-based information system.

This 'radial' model places the processor, the mechanism of data transformation, at the center of the representation. Historically, this and other elements decompose physically. The IBM 7094 mainframe, on which the first time-shared operating system was developed, expressed its central logic processor in multiple large cabinets. But for the purpose of this discussion, I will treat the central processor as atomic, a deliberate simplification that presents no discernable complications. (Multiprocessor systems typically present multiple processors as a single virtual element.)

The central processor is supported by elements within each of four arcs of functionality: working store, mass storage, peripheral/interface, and power and control.

4.1.1 Working Store

Working store is the locus of instructions and data composing the currently executing computation. While working store appears simple in modern processors (multilevel caching notwithstanding, as it is often opaque to other elements), working store in vintage systems can be quite involved, with considerations for segmentation and even differing behavior depending on segment. In systems involving slower memory technologies such as magnetic core, indirect addressing (a common scheme using the contents of a storage location as the address for the ultimate operand) often invoked a special flow of control within the central processor. Magnetic core is a persistent medium, while its successor, semiconductor memory, loses its content when power is removed.

4.1.2 Mass Storage (aka backing store)

Mass storage is where data and instructions are persisted, most importantly between operational phases (i.e., when the machine is turned off). Note that for early systems, mass storage was nonexistent or external to the electronic components (e.g., punched cards). The contents of mass storage are the primary domain of preservation through migration, and a key challenge is faithfully migrating bits from vintage to modern media. There are several levels of meaning in the representation of data on mass storage, often involving translations across the analogue/digital boundary.

4.1.3 Peripheral/Interface

Peripherals are the mechanisms by which the computation interacts with either human agents or other information systems or sources. This encompasses what we most often consider ‘human interface’ as well as data acquisition, process control and computer-to-computer connectivity.

4.1.4 Power and Control

The final arc, power and control, might seem to overlap with user interface but in fact is limited to those aspects of interaction that enable the system, not those related to the outcome of a computation. For example, an IBM 360 has an impressive front panel of lights and switches that are commonly used only for maintenance. Power systems can also be quite complex and are often problematic in vintage hardware.

4.2 Structure and substructure of elements

This model exposes a fractal nature. For example, a mass storage subsystem often encompasses considerable logic processing, sometimes to the level of a specialized processor for its control and data functions. Such subsystems interact with the central processor through a protocol or language at some level. As will be discussed shortly, it may be beneficial to consider the subsystem as a “black box”, or to decompose it further.

Even the central processor can exhibit this fractal nature: consider the microcoded processor, the predominant model since the 1970s. In this architectural model, a working store contains instructions that direct logic elements to perform the tasks necessary to implement the higher level, more abstracted model of the central processor. A microcoded processor includes most elements described above, with its peripheral arc presents the functionality of the processor to a higher level.

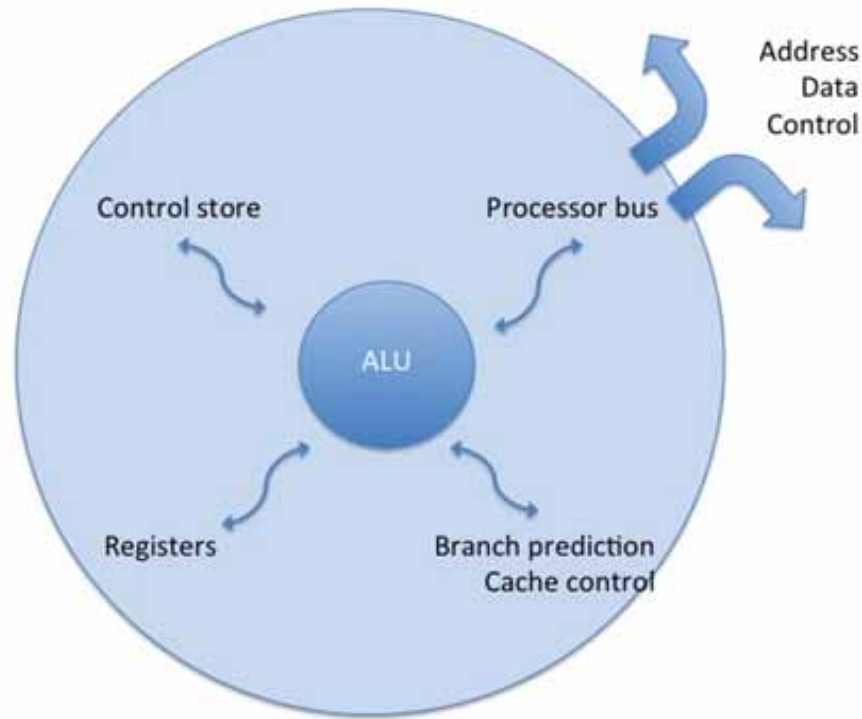


Figure 2. Fractal decomposition: a microcoded central processor.

4.3 Example: Massbus Disk Emulator

4.3.1 Background

LCM maintains a Digital Equipment Corporation DECsystem-20 Model 2065 system as a running machine, as near to full-time availability as possible. Among the elements of the running LCM system, the mass storage device for this system is commonly the RP06 removable-media disk drive. The media is a multi-platter disk pack that is inserted into a cavity in the drive unit, a cabinet about the weight and dimensions of a family-size washing machine.

Once the lid is closed, the pack spins up to a rotation speed of 3,500 rpm. Magnetic read/write heads moves across the platters of the pack to access concentric tracks of data. In order to avoid contact with the platters—which results in the mutual destruction of the head and the platter—they are designed to literally fly above the platter, employing the Bernoulli principle to maintain a distance less than the diameter of a human hair.

One of the weaknesses of this device is that human hairs—or other contaminants such as dust or smoke particles—can enter the cavity while it is open. When such detritus comes between the head and the platter, it often causes a “head crash”, a catastrophic event for both components, resulting in loss of both data and media.

The head crash that inspired the creation of the Massbus Disk Emulator (MDE) was not caused by a contaminant, but rather by a mechanical failure. Evidence suggests that the drive belt to the spindle driving the platter’s rotation became distorted and introduced oscillation in the spindle. When the team returned from a long weekend, the disk drive indicated an error and the media cavity was coated with a

fine layer of silvery powder (the platters are constructed of aluminum coated with ferrous material). Two disk heads were never found.

4.3.2 Analysis

Realizing that the number of RP06 disk units in existence, as well as the number of media packs, was growing only smaller, it was apparent that a new strategy was necessary if the 2065 was to remain operational on a regular basis. The decision was made to create an emulation of the RP06 disk drive.

Decomposing the functional arc represented by this disk drive unit, there were several choices that could have been made. The arc consists of an interface unit to the central processor, electrical media between the interface and the disk unit. The disk unit is composed of an interface to the electrical media, an interface to the analogue signals of the magnetic disk, and intermediary electronics that translates raw bits into data formatted to a particular specification, known as Massbus.

The following choices are feasible:

- The rotating disk media and analogue heads could be replaced by an analogue system that replicates the signals to and from the media.
- The digitized signals extracted from the disk could be replaced within the drive with digital data from some other sort of mass storage, with data written to the drive likewise translated through the drive's interface.
- The entire disk drive could be emulated, generating Massbus data and control signals to the connecting electrical media, and likewise interpreting received Massbus signals.
- The interface at the central processor could be replaced with an emulation that responds to bus signals directed toward the mass storage device.

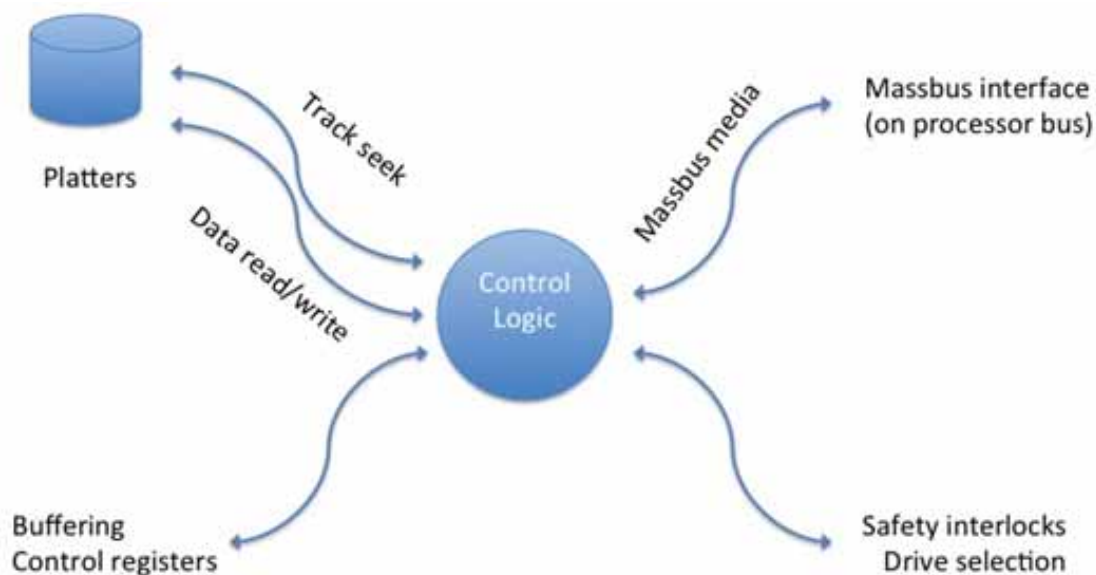


Figure 3. Decomposition of the mass storage arc for an RP06 disk drive.

All of these choices have been implemented for various systems; in particular, third-party vendors implemented the last idea in many variations.²¹ The LCM team chose to build an emulation of the entire RP06 drive at its juncture to the Massbus media (connecting cable). There were two particular arguments for this approach. For one, Massbus mass storage devices are employed by other DEC computers (PDP-10 in its various models, PDP-11 and VAX-11), thus potentially allowing the use of this one emulator with multiple systems. For the other, in generating the signals for a Massbus interface one can focus solely on one set of functionality, one that has been discussed in extant literature, while a bus-level connection to the central processor would require addressing many different data (even if only to ignore them).

4.3.3 Implementation

State machines were implemented with microcontrollers to present the data and control aspects of the Massbus disk. Source documents for the protocol were the technical manuals for the RH-11 and RH-20 interfaces for Massbus to the computer system bus. Four microcontrollers presented the Massbus interfaces and managed the Ethernet interface to the host data store.

Since the DECsystem-20 system requires system disk connectivity with both the PDP-11/40 boot processor and the KL-10 processor, the MDE was designed to copy the RP06 dual-port feature. The Massbus protocol allows up to eight devices to be connected to a single Massbus interface, so the MDE was intended to emulate eight Massbus disks on a single device.

4.3.4 Strengths and Weaknesses of the Approach

The greatest challenge in the design of the MDE was the availability and quality of documentation for the Massbus protocol. Although there was a formal specification for Massbus published by Digital,²² it was not known to the LCM team until after the first version of the MDE design was complete. According to a former highly-placed DEC employee, “[The standard] was never published at all outside the company and not widely consulted inside the company.”²³

The sources employed, official Digital documentation of two different processor-side Massbus interfaces,^{24, 25} were found in practice insufficient to describe the actual implementation. The experience of the LCM team strongly suggests that the Massbus Disk Emulator could not have been created solely from this seemingly authoritative documentation. The true “primary source” was a running system with a functional Massbus disk device.

Robert Supnik relates a similar experience in his development of the SIMH-based emulations that include Massbus devices,²⁶ and a darkly humorous take on these challenges is presented in “Tony in

²¹ The motivation of companies such as Emulex and Dilog was primarily to provide access to less expensive storage devices produced by companies other than the original equipment manufacturer.

²² Vic Ku, John Levy, and Pete Mclean, “Standard Mass Storage Interface--preliminary, Mass Storage, Interface Standard, and Massbus Interface Standard,” *DEC STD 159 Massbus Specification*, 1973.

²³ Private email exchange with Robert Supnik, June 14, 2012.

²⁴ Digital Equipment Corporation, “RH11-AB Option Description,” 1979.

²⁵ Digital Equipment Corporation, “RH20 Field Maintenance Print Set,” 1981.

²⁶ Bob Supnik, “A Massbus Mystery, or, Why Primary Sources Matter, Even In Computer History,” 2004, <http://simh.trailing-edge.com/docs/massbusmystery.pdf>.

RH20 Land.”²⁷ These experiences should give pause to those who suggest that future practitioners will be able to create Rosetta stones for digital documents from specifications and protocol documents.

Once constructed, the MDE has proven to be very reliable and is routinely used for storage of user account data on LCM’s running 2065 system. As it has matured, its failures have been limited to external issues with the Museum’s power grid, and the MDE fails “safely” compared to an RP06.²⁸ The supporting server that contains track data is constructed with modern RAID and other error-correcting mechanisms to avoid corruption of data. It is routinely backed up with modern tools that also include error detection and correction.

4.3.5 Summary: Facility for Access and Preservation

The Massbus Disk Emulator provides expansive and robust mass storage for vintage systems employing the Massbus protocol, enhancing access to large quantities of data on such systems. The MDE and its supporting server employ modern components that provide greater structural accessibility due to smaller physical and power footprints. Once the MDE has been proven for other systems that support Massbus, the feasibility of running these systems is improved. For example, while a KS-10 can be run on household power, the RP06 disk drive requires 240V with considerable surge current capability.

The MDE supports long-term preservation by providing that data can be housed on current and future mass storage systems that support error detection and correction as well as best-practice backup protocols. Its file formats on the modern host are well documented to ensure that migration can be accomplished as necessary with little risk of corruption and, in a worst case, recovery is not obfuscated. The meaning of the content exists at a low level of abstraction within the information system, requiring a minimal and straightforward amount of metadata to express and preserve it. The connection of the expressed functional arc component is clear, concise and distinguished.

4.4 Short Stories: Terminal Emulation

Teletypewriters (colloquially, Teletypes, after the name of the primary vendor) and video display terminals (VDTs) are historically significant elements of the peripheral functional arc. In the earliest days of what is broadly called interactive computing, Teletypes were appropriated from the telecommunications industry to provide textual interface with computer systems. VDTs were created as alternatives to Teletypes, as the latter were electromechanical devices that required constant maintenance, generated significant audible noise and were limited in speed of interaction.

Terminal emulators running on modern PCs provide inexpensive and widespread access to character-cell systems. They provide access to both directly connected and network-proxied systems by obviating the need to possess and maintain one of the dwindling number of vintage Teletypes or VDTs. They preserve the essential experience of interaction with a character-based terminal, although modern features (for example, a mouse) may detract from the experience of a pre-GUI interface.

²⁷ Anthony Wachs, “Tony in RH20 Land,” <http://www.inwap.com/pdp10/rh20.txt>.

²⁸ Since the MDE does not function sufficiently well with the PDP-11/40 boot processor, LCM’s 2065 still employs a real RP06 as the ‘system’ device. In order to protect the RP06, LCM inserted a drop-out relay in its power line. The power failures at the Museum are often ‘bouncing’ failures, and LCM has observed that such up-and-down power scenarios create mechanical issues as the disk head armature is intermittently powered.

4.5 Short Stories: Power Systems

Power supplies are a mundane Achilles' heel for information systems. Their failure can render a system inoperable: one hopes for benign failure but in the worst case, the failure of a power supply can destroy functional elements of the system.

As discussed above, a key weakness in power supplies is the use of aluminum electrolytic capacitors to filter rectified alternating current into direct current. In older systems, this is accomplished in *linear* power supplies. According to a major manufacturer of these components, electrolytic capacitors will fail over time, *regardless of whether they are in service*.²⁹ Their replacement is problematic, as new components are often dramatically different in form factor.

In some systems at LCM, original power supplies have been supplemented with modern *switching-mode* units. The new units are diminutive in size and do not substantially affect the visual impact of the original system. They are also dramatically more energy efficient. Through careful choices in mounting and connection, no holes are drilled and no wires are cut.

This will be of less importance for maintenance of running systems built with switching-mode supplies, but has been an efficacious strategy in working with systems from the 1960s and early 1970s.

4.6 Will the Center Hold? Emulating the Central Processor

One might infer that the central processor is considered the one element that must remain inviolate, but there are cases demonstrating the feasibility of emulating it as a distinct element of the information system. For example, when the PDP-10 processor line was discontinued by DEC, members of the Large Systems Group started a new company that ultimately created an FPGA-based emulation of the PDP-10 processor. Contesting the idea that the PDP-10 architecture could not scale *down* to become a desktop machine, a feat accomplished by the VAX, they sought to create a “Ten On A Desktop”. In the mid-1990s, XKL, Inc. offered the TOAD-1, a microcoded implementation of the PDP-10 architecture. This system was more of a *desk*side machine, but proved the argument regarding scale—too late, as 32-bit microprocessors now dominated this space. XKL continues to use their emulated PDP-10 processor as the core of high-end networking equipment.

4.7 Summary

Emulation of a component effectively reproduces all or part of the arc of functionality for that component. It provides clearly defined interfaces between elements that support a progressive approach to building a complete information system incorporating some percentage of the original system, perhaps approaching zero as time goes by. As each element's semantics is far less complex than that of the entire system, the decomposed system is more amenable to understanding and analysis.

²⁹ Sam G. Parler and Cornell Dubilier, “Reliability of CDE Aluminum Electrolytic Capacitors,” (n.d.): 1-10, <http://www.cde.com/tech/reliability.pdf>.

5. Conclusions

5.1 Where We and Our Turtles Stand

Digitization *per se* is a mechanism for preserving and providing access to pre-digital documents. Discussions regarding digitization often do not adequately address the issues for born-digital documents, especially those dynamic digital documents whose meaning can best, or perhaps *only* be expressed within the documents' original execution environment.

Preservation of the original execution environment serves both static and dynamic documents. With the former, it reduces data loss in the transcription process of migration. Regarding the latter, it is important to recognize that the term 'preservation' must be construed differently than is most often employed by archival institutions: it must also encompass the preservation of functionality.

The two most commonly referenced means of such functional preservation are often seen dialectically. In truth, there are dependencies that will exist for a period of time that is probably finite but not well bounded. For some time to come, the preservation of running systems will be necessary to provide primary-source documentation for software emulations and to validate their correctness. Otherwise, we can have no confidence that software emulations can serve long-term preservation of digital documents. The end of our dependency on running systems as primary source documents is not in sight.

5.2 Future Work

For the preservation of dynamic digital documents based on systems now considered historical, we must better understand the resolution curve of the correctness of emulations. Software Quality Assurance practice can be helpful, but preservation will be best served if preservationists adopt domain expertise regarding execution environments. Computer-based information systems cannot be black boxes to digital librarians. For our current history, our work is far from done.

We have the opportunity to learn from the challenges we face today in generating standards and best practices for digital documents in the future, to minimize the dependencies we see on platforms for documents in antiquity. A combination of migration and emulation may provide the greatest access, but we will need to continuously reassess the issues of preservation and archiving of working environments, even as we attempt to decouple future documents from their substrates. We will not find closure soon.

If we pursue these ideas naively and seek quick fixes, we will almost certainly continue to build tall towers of turtles, which are fragile both for the individual document—the potential for imperfect migration or poor interpretation of functionality—and for preservation in general, as we may find ourselves wasting valuable resources on maintaining dependency structures—feeding towers of turtles—that could have and should have been collapsed and obviated a long, long time ago.

Acknowledgements

I want to thank to the Living Computer Museum and Paul Allen, its principal, for allowing me to document the story of the Massbus Disk Emulator project for this paper, and in particular to Keith Perez for our discussions of the particulars of this work. Thanks also to Robert Supnik, the creator of SIMH, for our discussions of the development of the SIMH package, his approval of my use of the SIMH history in this paper, and our conversations of his experiences at Digital Equipment Corporation.

Challenges to Capture the Hybrid Heritage

When Activities Take Place in Both Digital and Non-Digital Environments

Erik Borglund

Mid Sweden University, erik.borglund@miun.se

Abstract

Today our life is integrated by various uses of IT. Since the 1990s archival science has struggled to establish new methods that support the management of records born digital. In this article we claim that there is still a distinct border between the digital and the non-digital and that modern archival methods do not deal with records that are “hybrids”, i.e., a mix of both digital and non-digital. Why is this problematic for a wide heritage approach? One problem with the current approach is that the records captured are not fully representing all the activities they encompass. It is not possible to understand the full range of the individual’s behaviour because some activities have not been captured. In this article two narratives are used. These narratives will exemplify the problems that come from modern records management in a hybrid environment.

Author

Erik Borglund, Ph.D. in computer & system science, is a senior lecturer and researcher at the Archival and Information Management School of Mid Sweden University, and his main research interests are digital recordkeeping, document management, information systems in crisis management, information systems design and Computer Supported Cooperative Work (CSCW). Erik Borglund was a sworn police officer for 20 years, before he became a 100% academic.

1. Introduction

How can the full extent of people’s activities be captured in modern environments? This is the topic for this article. The article discusses how the current methodologies for records capture and records management are failing to provide a true and authentic picture of our current lives. But first we take a short retrospective journey to very briefly summarize some of the challenges that the archival research community has been struggling with over the last 20 years.

At least during the last 20 years it has been well known that information technology has changed our society and this is also the case for archival science as well as the more practical work of archivists and records managers. Today the concept of digital records is a very familiar and known phenomena as digital records have been common in modern administrative environments since the early 1990s. Nearly 20 years ago the first researchers started a lively debate around the problems related to computerization, and the fact that more and more records were born digital.¹ One identified problem was that techniques

¹See David Bearman, *Electronic Evidence : Strategies for Managing Records in Contemporary Organizations* (Pittsburgh: Archives and Museum Informatics, 1994); David Bearman, “Record-Keeping Systems,” *Archivaria* 36 (1993): 16-36; Charles M. Dollar, *Archival Theory and Information Technologies : The Impact of Information Technologies on Archival Principles and Methods*. Informatics and Documentation, series 1 (Macerata, Italy: Univ. of Macerata, 1992); Trevor Liverton, *Archival Theory, Records, and the Public* (Lanham, MD; London: The Society of American Archivists and Scarecrow Press, 1996); David Roberts, “Defining Electronic Records, Documents and Data,” *Archives and Manuscripts* 22, no. 1 (1994): 14-26.

and methods available for management of records that were developed for an analogue environment are not best suited for a computer-based environment were questioned. For example, the Swedish archival practice was almost 100 years old when the problem with digital records was raised. In the literature you can find a pattern. First the main focus within archival science was to focus on what was maybe the most acute problem, i.e., how to preserve the digital records.² This problem seemed to be so extensive that some researchers also proposed that the digital records was the birth of a new archival paradigm.³

Just as the archival community started to learn how to manage digitally born records the use of Internet in and between individuals and organizations had come to the point where we can say that the E-society started. Today E-services, E-governance, E-government, E-democracy are so accepted in society that the prefix “E” almost has had its day. However the E-services and especially the fully integrated e-services brings new problems for the archival community. One of the many challenges is that in a fully integrated e-service records are created in an environment owned and controlled by many actors, that put current records management theories on test.⁴

This short introduction aims to argue for the fact that the technical development will continue and it will be creating new challenges for the archival community, and the possibility to preserve digital records over time.

All records, in both electronic and traditional formats, are evidence of business in organizations.⁵ To maintain the evidential value of a record it is necessary to preserve the content of the record, to

² See e.g., Bearman, *Electronic Evidence*; Bearman, “Record-Keeping Systems”; Dollar, *Archival Theory and Information Technologies*; Charles M. Dollar, *Authentic Electronic Records : Strategies for Long-Term Access* (Chicago, IL.: Cohasset Associates, 2000); Luciana Duranti, “Concepts, Principles, and Methods for the Management of Electronic Records,” *The Information Society* 17 (2001): 271-79; Luciana Duranti, “The Impact of Digital Technology on Archival Science,” *Archival Science* 1, no. 1 (2001): 39-55; Luciana Duranti, “The Impact of Technological Change on Archival Theory” (paper presented at the International Council on Archives Conference, September 16, 2000, Seville, Spain); Anne J. Gilliland-Swetland and Philip B. Eppard, “Preserving the Authenticity of Contingent Digital Objects,” *D-Lib Magazine* 6, no. 7/8 (2000), <http://www.dlib.org>; Anne J. Gilliland-Swetland, *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment* (Washington, D.C.: Council on Library and Information Resources, 2000); Sally McInnes, “Electronic Records: The New Archival Frontier,” *Journal of the Society of Archivists* 19, no. 2 (1998): 211-20; Frank Upward, “Modeling the Continuum as Paradigm Shift in Recordkeeping and Archiving Processes, and Beyond - a Personal Reflection,” *Records Management Journal* 10, no. 3 (2000): 115-39.

³ Terry Cook, “What Is Past Is Prologue: A History of Archival Ideas since 1898, and the Future Paradigm Shift,” *Archivaria* 43 (1997): 17-63; Bruno Delmas, “Archival Science Facing the Information Society,” *Archival Science* 1 (2001): 25-37; Gilliland-Swetland, *Enduring Paradigm*; Upward, “Modeling the Continuum.”

⁴ See e.g., Karen Anderson, Erik A. M. Borglund, and Göran Samuelsson, “Strategies for High Quality Recordkeeping in Public E-Services,” *iRMA - Information & Records Management Annual 2010* (2010): 73-82; Viveca Asproth, Erik A. M. Borglund, Göran Samuelsson, and Lena-Maria Öberg, “E-Tjänstens Framtida Historia: Informationsbevarande, Ett Bortglömt Ansvarsområde?” in *Förvaltning Och Medborgarskap I Förändring*, ed. K. Lindblad Gidlund, A. Ekelin, S. Eriksén and A. Ranerup (Lund: Studentlitteratur, 2010), 167-84; Erik A. M. Borglund, Karen Anderson, and Göran Samuelsson, “How Requirements of Record Managers Change after Implementing New Electronic Records Management Systems,” in *3rd European Conference on Information Management and Evaluation, Gothenburg, Sweden, 2009*; Göran Samuelsson and Lena-Maria Öberg, “The Future History of E-Services. Long-Term Preservation of Complex and Integrated E-Services,” in *Collaboration and the Knowledge Economy: Issues, Applications, Case Studies. Part I, 2008*, pp. 286-293; Göran Samuelsson, Lena-Maria Öberg, and Erik Borglund, “E-Services and Long-Term Preservation,” in *E-gov 07 conference, Regensburg (Germany), September 3-7, 2007*.

⁵ See Duranti, “Concepts, Principles, and Methods”; National Archives of Australia, “Digital Recordkeeping: Guidelines for Creating, Managing and Preserving Digital Records,” ed. National Archives of Australia, 2004;

preserve the context where the record was created, and to preserve the structure of the record.⁶ In 2004 The National Archives of Australia listed common places where records can be created.⁷ By using office applications records can be found in word processed documents, desktop processed documents, spreadsheets, and in presentations; Records can be born using different business information systems such as databases, geospatial information systems, human resource systems, financial systems, workflow systems, client management systems, and in customer relationship management systems; Records can also be born and generated in different web-based environments such as intranets, extranets, public websites, and in online transactions; Records can also be the result of different communication technologies such as email, SMS, MMS, electronic fax, voice mail, and instant messaging. Today this list is a reality, and not something strange at all.

In 2004 the web 2.0 came and that resulted in social media,⁸ and in 2006 the concept of enterprise 2.0 entered the scene, which brought the use of social media into companies as well as organizations. We all know what happened after the birth of the iPhone and the Google Android OS, a new way of becoming on-line almost every day began. Suddenly everyone knew what an “app” is, and programming competitions was renamed as “appenings”. Today the use of Internet, social media through smart phones and tablets are widespread, and we live our lives more online than ever.⁹

This article will aim to present potential problems related to preserving the full extent of the modern human activities, and by that capture the future of our history. The article departs from that our lives becomes more and more integrated by various uses of every day IT, web technologies as, for example, social media, and virtual communities. What kind of yet not identified challenges can be identified from the new and modern use of IT?

This article is structured as follows. First a theoretical section is found in relevant theories are presented. The theoretical section follows by a method section and then a result section. The article ends with a final conclusion.

1.2 Applicable Theoretical Approaches

When more and more activities are taking place on the Internet, and we all can be online even when we are away from our computers by using smart phones as well as tablets. Today there are many people that argue for that the “young generation” is digital natives. Digital native is a term that is used about humans that are born during or after the wide spread introduction of IT in modern society. The idea with a digital native concept is that the borders between virtual or digital activities and real life activities are vague and the digital natives act as there are only one “world”.¹⁰ However modern research also indicate that the

Barbara Reed, “Records,” in *Archives: Recordkeeping in Society*, ed. Sue McKemmish, Michael Piggott, Barbara Reed, and Frank Upward (Wagga Wagga: Charles Sturt University, Centre for Information Studies, 2005), 101-30.

⁶ National Archives of Australia, “Digital Recordkeeping”; Reed, “Records”; Theo Thomassen, “A First Introduction to Archival Science,” *Archival Science* 1, no. 4 (2001): 373-85.

⁷ National Archives of Australia. “Digital Recordkeeping,” 13.

⁸ Tim O’Reilly, “What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software,” *Communications & strategies* 65 (2007): 17-37.

⁹ Lenhart Amanda, Kristen Purcell, Aaron Smith, and Kathryn Zickuhr, *Social Media & Mobile Internet Use among Teens and Young Adults* (Washington, D.C: Pew Research Center, 2010).

¹⁰ The term is further elaborated by Marc Prensk, “Digital Natives, Digital Immigrants Part 1,” *On the Horizon* 9, no. 5 (2001): 1-6; and Marc Prensk, “Digital Natives, Digital Immigrants Part 2: Do They Really Think Differently?” *On the Horizon* 9, no. 6 (2001): 1-6.

term digital natives have been criticized because it is problematic to identify one generation or type of humans that actually adopt IT in a very advanced way.¹¹

Information systems research has a long tradition in studying the interaction between users and IT. Throughout the last 15 years there are several examples on contributions where various approaches have been taken to study IT in organizations and to understand how IT is used in organizations.¹² One of the ways to understand the complex relationship between IT, organization and user is by focusing on how agents act in for example an organization, thus using the concept agency. In a debate section in the *Scandinavian Journal of Information Systems* in 2005 Rose *et al.*¹³ debated on the problem of agency, in which several perspectives on how to study IT and organizations using the concept of agency was presented. Two of the most frequently applied theories in IS research were put against each other, the *Structuration theory* and the *Actor network theory* (ANT), which open up different ways to interpret agency. As a complement to using either of the above theories to explain the relationship between IT and organizations and the users, where the user are often seen as separated from technology. One can treat the organization/user and IT as one unit as, for example, a hybrid. Mike Michael¹⁴ has developed the hybrid perspective derived from ANT, and Lindroth¹⁵ takes Michael's hybrid perspective and creates the "laptop", which is a laptop and its user seen as a single unit. Lindroth¹⁶ uses this hybrid as unit of analysis in a larger ethnographic study, where he uses the hybrid in form of the laptop to understand agency. The hybrid perspective also conforms to the concept of sociomaterialism, where, for example, Orlikowski argues that "all practices are always and everywhere sociomaterial."¹⁷

1.3 The problem further described

In this article we will apply the hybrid perspective, but even if we assume that more and more individuals are digital natives, and no longer digital immigrants, there remain very distinct boundaries between the digital and the non-digital. As presented in the introduction the archival science has struggled to establish

¹¹ S. Bennett and K. Maton, "Beyond the 'Digital Natives' Debate: Towards a More Nuanced Understanding of Students' Technology Experiences," *Journal of Computer assisted learning* 26, no. 5 (2010): 321-31.

¹² Amongst many contributions see Rob Kling, "Learning About Information Technologies and Social Change," *The Information Society* 16, no. 3 (2000): 217-32; Rob Kling and Roberta Lamb, "IT and Organizational Change in Digital Economies," *Computers and Society* 29, no. 3 (1999): 17-25; Roberta Lamb, "Social Actor Modelling in Ict Research" (paper presented at the 12th European Conference on Information Technology Evaluation (ECITE), Turku, Finland September 29-30, 2005); Roberta Lamb and Rob Kling, "Reconceptualizing Users as Social Actors in Information Systems Research," *MIS Quarterly* 27, no. 2 (2003): 197-235; Wanda J. Orlikowski and Jack J. Baroudi, "Studying Information Technology in Organizations: Research Approaches and Assumptions," *Information Systems Research* 2, no. 1 (1991): 1-28; Wanda J. Orlikowski, "The Duality of Technology: Rethinking the Concept of Technology in Organizations," *Organization Science* 3, no. 3 (1992): 398-427; Wanda J. Orlikowski, "Sociomaterial Practices: Exploring Technology at Work," *Organization Studies* 28, no. 9 (2007): 1435-48; Wanda J. Orlikowski and Debra C. Gash, "Technological Frames: Making Sense of Information Technology in Organizations," *ACM Transactions on Information Systems* 12, no. 2 (1994): 174-207.

¹³ Jeremy Rose, Matthew Jones, and Duane Truex, "Socio-Theoretical Accounts of IS: The Problem of Agency," *Scandinavian Journal of Information Systems* 17, no. 1 (2005): 133-53.

¹⁴ Mike Michael, *Reconnecting Culture, Technology, and Nature: From Society to Heterogeneity* (International Library of Sociology, London: Routledge, 2000); Mike Michael, *Technoscience and Everyday Life: The Complex Simplicities of the Mundane* (Maidenhead: Open University Press, 2006).

¹⁵ Tomas Lindroth, "The Laptop: Understanding Agency from a Hybrid Perspective," in review for the *Scandinavian Journal of Information Systems* (forthcoming).

¹⁶ Ibid.

¹⁷ Orlikowski, "Sociomaterial Practices."

new methods and technologies that support the management, i.e., creation, capturing, and preservation of records born digital, since the early 1990's. These technologies can be argued to focus on pure digital records, i.e., records resulting from activities that in some way are recorded by digital technology. But if one scrutinize how modern individuals act, one will see many examples of activities that take place in both digital and non-digital environments. For example: A project organization may start with an online meeting Google+ Hangout, where several project leaders discuss the new project and the current progress. After the meeting, each project leader go back to its own project staff, and directly sets up a live meeting where they carry out an information modeling exercise with sticky notes. In this example, activities are taking place in two co-existing worlds, but the involved actors do not act as if there were boundaries between these two worlds.

Why is this problematic for a wide heritage approach? One problem with the current approach is that the records captured are not fully representing all the activities that produce them or they refer to. It is not possible to understand the full range of an individual's behaviour because some activities have not been captured. Activities that take place in modern environments are both digital and non-digital. When records are artificially aggregated without assurance that they are the complete output of an activity, authenticity and reliability of the records can then be challenged.

2. Method

In this article two narratives will be used to present the problem and exemplify how the problem look like when one want to capture the heritage in a modern environment. The first narrative is based upon a research that have focused upon how modern criminals act as hybrids and what problem that gave the police. Even if the narrative is based upon a case which is far away from the heritage domain it will be usable to discuss the problems related to capture the hybrid heritage. The second narrative is a story from the authors everyday work as senior lecture at a Swedish University in which both online activities with students takes place parallel as offline activities take place. Sometimes they take place simultaneously.

The research that is the basis for the first narrative was a qualitative case study, and the research have been published in various contexts.¹⁸

The second narrative is created using what best is described as a autoethnographical metod,¹⁹ i.e., where experience of the author is used to present a story on which one can reflect. However the method in this article is only for building up a narrative that can be used for further discussions.

¹⁸ Erik A. M. Borglund, Tomas Persson Slumpi, and Lena-Maria Öberg, "A Success Story About the Investigation of Organized Prostitution: The Jämtland Police Authority Case," in *Third Nordic Police Research Seminar* (Umeå, 2010); Erik A. M. Borglund, Lena-Maria Öberg, and Thomas Persson Slumpi, "Hybrids Acting at the Hybrid Arena – Investigating Crimes Committed by Digital Natives," in *Selected Papers of the Information Systems Research Seminar in Scandinavia Nr. 2 (2011): Iris 34 Ict of Culture – Culture of Ict*, ed. Judith Molka-Danielsen and Kai Kimppa (Tapir Academic Press, 2012); Erik A. M. Borglund, Lena-Maria Öberg, and Thomas Persson Slumpi, "Success Factors for Police Investigations in a Hybrid Environment: The Jämtland Police Authority Case," *International Journal of Police Science and Management* 14, no. 1 (2012): 83-93; Erik A. M. Borglund, Lena-Maria Öberg, and Thomas Persson-Slumi, "Hybrids Acting at the Hybrid Arena" (paper presented at the IRIS 2011, Turku, 2011).

¹⁹ Carolyn Ellis, Tony E. Adams, and Arthur P. Bochner, "Autoethnography: An Overview," *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 12, no. 1 (2011); Nicholas L. Holt, "Representation, Legitimation, and Autoethnography: An Autoethnographic Writing Story," *International Journal of Qualitative Methods* 2, no. 1 (2003): 18-28.

These two narratives will then be used as basis in a logical discussion and analysis of common problems and challenges upon how to capture the hybrid heritage.

3. Results

In this section two narratives where the concept of Hybrid will be highlighted and presented. The first narrative is based upon a real large police investigation about gross procuring. In this narrative the male pimp is a hybrid, and his activities will serve as basis for further discussion on the problems related to be able to capture the hybrid heritage.

3.1 Narrative 1

A 40-year-old male, in the following text called *the Pimp*, was the organizer of a large prostitution business. He had five women working for him, one of which one was his wife. The Pimp was responsible for the advertisement of the sexual services the five women offered. To advertise the services the Pimp used Internet but also the teletext (Text TV). There are several sites available for this purpose, and the communication and advertisement were done openly. In the advertisements the potential buyers were required to contact the women by mail, SMS or a phone call. The Pimp provided the women with mobile telephones so that they could communicate with potential customers. The women were not stationed in one place but travelled between several cities mostly in the mid of Sweden, for example the counties of Västernorrland and Jämtland. Occasionally they also travelled to other cities in the more southern parts of Sweden. The Pimp arranged the travels as well as reserving and paying for hotel rooms, etc. It was no question that he controlled these women and used them in his business.

A surveillance team in a Swedish police authority investigated the case, and the team consisted of both Internet experts as well as traditional street officers. A criminal intelligence analyst that worked at a regional intelligence office did the Internet analysis, and she was also the person that triggered the whole investigation. One part of the analysts work was to scan the Internet for web pages and web forums at which sexual services were offered, hence potential violations of the law that prohibits purchase of sexual services. The analyst noticed that there were a couple of women that seemed to actively work as prostitutes mostly in the mid Sweden region. The women seemed to be like prostitute nomads, who moved their business around. The women's activities could be mapped into a cluster of activities with the Centre in Västernorrland County. The analyst could through traces left on Internet map patterns of how the women moved, and how they communicated their presence in advance.

In the formal police investigation the police needed to use both the digital traces found on Internet about the male pimps activities, and more traditional evidences about the pimps activities in the non-digital world. The police needed to put together a chain of evidence that could be presented as a whole, and in which the activities online as well as offline could be presented in chronological order.

The police managed to provide the public prosecutor enough material for public prosecution, and the pimp was sent to jail for a long time.

3.2 Narrative 2

Archival education at Mid Sweden University is carried out in a blended learning environment, which means that you as student can both follow the lectures at campus as online. The standard course is 10

weeks long and starts with an intensive three day lecturer where the students either visit our university campus or participate online. The students can also choose to not participate at all, and only listen to the recordings afterwards.

We use Adobe® Connect™ as video conference tool, and that is the tool we use to enable students to follow lectures that take place at campus. The Adobe® Connect™ is an Internet based service that you run in your ordinary web browser and that make it possible to share screen, interact with voice and text, and with a web cam. At the campus class room the student do not see what is going on Adobe® Connect™, they only see the lecture slides on a SMART Board™ The lecture slides is in Adobe® Connect™ shared with the online students.

As all modern universities we use a learning management system (LMS) for managing the course content, lecture notes, and everyday communication with the students. In the LMS we provide static material as reading instructions, literature lists, literature for download. We also provide materials from the lectures in form of lecture notes, lecture slides, recordings. In the LMS there is also more dynamic part in which we as teachers can communicate with the students both in a discussion forum but also as more direct messages. We can also chat with students that are logged in at the same time as us in the LMS.

Beside the LMS, the students might communicate with me as teacher through the ordinary e-mail system or comment on the blog that I write as researcher on the University Web page.

3.3 Summary of the two narratives

The two narratives aims to present the fact that today the activities a person is carrying out is both done in the “non-digital world” and in the digital world, foremost on Internet, and these activities can be done and carried out simultaneously. In the digital world digital footprints are left, i.e., digital traces left in the digital environment, and they together can form the evidence of an activity. In the “non-digital world” the activities needs are leaving other kind of traces and evidence.

4. Discussion

In this section we will use the above presented narratives as basis for presenting the problems found related to preserving the hybrid heritage as we call it in this article.

First we must characterize the hybrid. The hybrid is the actor that is using information technology in a way that makes the IT almost as a part of the actor. One example that many are familiar with is all the people you see interacting with their smartphones in all kind of situations. The smartphone can be seen as an extension of the body, connected to Internet. It make it impossible to separate the use of IT from the use and acting as a human.

In the above first narrative the pimp is another hybrid. He acted as a pimp on Internet, i.e., in the “digital world”, and in the “physical world”. He did not use IT as mean to commit a crime, he used IT as part of his action and we argue that he can be seen as a hybrid. In the same way is it possible to interpret and see the teacher role as a hybrid in the last narrative. When we act as a teacher we interact with our students simultaneously in the “physical world” and in the “digital world”. This article was not aiming to fully elaborate and theoretical argue for how to motivate and argue that these examples really are hybrids. Based upon the work by Michael²⁰ and Lindroth²¹ these examples are possible identify as hybrids.

²⁰ Michael, *Reconnecting Culture*; Michael, *Technoscience*.

The research behind the first narrative resulted in a concept we call “hybrid arena” which is the environment in which the hybrid act. The hybrid arena is with archival terminology possible to be understood as the full context in which the hybrid act.

4.1 Discussion about the problem

If one look at the hybrid presentation above and discuss whether the current concepts of how to manage records a series of problems arises.

If you apply the current approach to manage records in a hybrid environment you realize that the records that are captured are only representing the activities in one of the “worlds” either the physical world or either the digital world. This means that the records captured are not fully representing all the activities that produce them or they refer to. For example to fully understand how we as teachers interact with both the campus students and the online students in the blended learning environment is very complicated if the records captured are not covering the two environments. It is maybe even more clear in the first narrative where the activities by the police is separated by two actors one the intelligence analyst and the traditional surveillance police officers. Their activities are recorded individually and the activities do not correspond with the activities the pimp does. He is moving between the two worlds, the digital world (Internet) and the non-digital world as a hybrid. When the police try to record and document his activities they cannot capture it as a coherent story.

Activities that take place in modern society are both digital and non-digital. When records are artificially aggregated without assurance that they are the complete output of an activity, authenticity and reliability of the records can then be challenged.

There are probably many that do not agree with the problem the hybridization will result in for capturing records of modern activities. But one of the new main arguments for this problem is that the hybrid is a new way to understand use and interaction with technology. To capture the hybrids activities in a hybrid environment, new tools are needed. The current tools focus on records capture in one environment solely.

5. Concluding remarks

This article will aim to present potential problems related to preserving the full extent of the modern human activities, and by that capture the future of our history.

Current approaches to capture records are in this article argued to be less optimal for capturing the modern activities that take place in both the digital and the non-digital world. One problem with the current approach is that the records captured are not fully representing all the activities that produces and encompasses. It is not possible to understand the full range of the individual’s behaviour because some activities have not been captured. Activities that take place in modern environments are both digital and non-digital. When records are artificially aggregated without assurance that they are complete, authenticity and reliability of the records can then be challenged. In this article two narratives were used to exemplify how the modern hybrid actor could behave and the problems presented was related to those narratives.

²¹ Lindroth, “The Laptoper.”

Limited Resources or Expertise: Case Studies in Addressing the Issue

La problématique de la préservation de la mémoire collective au Burundi à l'ère des NTIC¹

Étude de cas menée à la Cour suprême

Jean Bosco Ntungirimana

Vice Président de l'Abadu-Burundi

Résumé

Les archives burundaises se trouvent dans une situation alarmante. L'état des documents produits ou reçus est inquiétant dans plusieurs structures, car celles-ci font souvent face à de multiples problèmes liés : au manque d'équipements adéquats; au manque de personnel qualifié; à l'impossibilité de repérer l'information sensée existante; à la difficulté de maîtriser l'inflation documentaire; à l'absence d'une législation archivistique; à l'exiguïté des locaux; etc. Une étude fut donc menée du 24 au 27 avril 2011 avec but d'identifier ces problèmes pour ensuite exécuter la mise en œuvre de programmes efficaces quant au développement d'infrastructures modernes, et ce, pour pouvoir permettre : aux citoyens d'accéder rapidement à l'information recherchée; l'adoption et l'harmonisation d'une législation sur la gestion électronique des documents; et la gestion efficace des documents électroniques pour garantir la fiabilité de l'information. Cette communication s'appuie sur les résultats de l'étude susmentionnée et sur une série des recommandations basées sur les conclusions qui ont été formulées et transmises aux gouvernements respectifs de la Communauté d'Afrique de l'Est (EAC).

Auteur

Jean Bosco Ntungirimana a obtenu sa licence en sciences de l'information documentaire, avec une spécialisation en archivistique de l'École de bibliothécaires, archivistes et documentalistes de Dakar. Il travaille en tant que consultant des archives et bibliothèques et est auteur de divers articles et études touchant sur les archives, la gestion des dossiers et les bibliothèques. Depuis 2004, il est directeur du département de la documentation et des archives à la Cour nationale des comptes. Il est membre de diverses associations professionnelles à l'échelle nationale et internationale, incluse parmi elles le conseil international des archives (ICA), l'*International Federation of Library Associations and Institutions* (IFLA) et l'Association Internationale francophone des bibliothécaires et documentalistes (AIFBD). Il siège aussi au sein du comité exécutif du *Burundian's Association of Librarians, Archivists and Documentalists*, en tant que vice-président et membre fondateur.

1. Introduction

Du 24 mars au 27 avril 2011, grâce au financement octroyé par le Centre de recherches pour le développement international (CRDI), nous avons mené une étude de recherche commanditée par l'*International Records Management Trust* et le secrétariat de la Communauté d'Afrique de l'Est (EAC). Le thème de cette étude est : « Recherche CRDI — Aligner la gestion d'archivage avec les technologies de l'information et de la communication, l'e-gouvernement,² et de la liberté d'accès aux informations dans la communauté de l'Afrique de l'Est : Cas de la Cour suprême du Burundi ». Étant donné que les dossiers constituent les fondements de la justice et fournissent les preuves nécessaires pour soutenir

¹ Nouvelles technologies de l'information et de la communication (NTIC).

² Aussi connu sous l'appellation « cybergouvernement » (CAN) ou « administration électronique » (FRA, EU).

l'autorité de la loi, protéger les droits, améliorer les services aux citoyens, gérer les ressources, et encourager les stratégies de responsabilité et de lutte contre la corruption, cette étude a pu identifier un certain nombre de problèmes et de lacunes fondamentaux pour la mise en œuvre de programmes efficaces de préservation du patrimoine documentaire dans l'environnement électronique.

Au cours des entretiens menés auprès de certains membres du personnel de la Cour suprême et de certains justiciables rencontrés sur place, on constate que les archives sont d'une grande importance dans le fonctionnement de l'appareil judiciaire en général et, plus particulièrement, de la Cour suprême. Étant la plus haute juridiction ordinaire de la République du Burundi, la Cour suprême reçoit une grande masse de dossiers provenant des juridictions inférieures, tels que les cours et tribunaux (par ex. : les tribunaux de grande instance et les cours d'appel). C'est pour cette raison que les dossiers de la Cour suprême revêtent un caractère particulier et méritent une attention particulière lors de leur conservation. Toutes les personnalités rencontrées ont émis le souhait de voir les archives de cette cour réorganiser afin d'aider les justiciables à retrouver facilement leurs dossiers, car on a remarqué des files d'attente des justiciables venus demander l'état d'avancement de leurs dossiers. Les greffiers nous ont confirmé qu'ils reçoivent au moins deux cent cinquante dossiers par mois, ce qui cause un casse-tête dans leur gestion quotidienne d'autant plus que la Cour suprême ne dispose pas en son sein un espace suffisant pour la conservation optimale des dossiers réceptionnés et des professionnels d'archives chargés de gérer tous ces documents.

Malgré cette étude, il n'est pas possible de maintenir un système judiciaire efficace et juste sans tenir des dossiers précis sur la conduite des affaires individuelles, et sans pouvoir accéder aux informations dans ces dossiers quand elles sont nécessaires. De plus, les résultats de cette étude montrent que la Cour suprême du Burundi dispose d'une grande masse de documents d'archives (dossiers d'affaires) et nécessite d'abord d'amélioration dans le domaine de la technologie de l'information et de la communication (TIC), l'e-gouvernement et l'accès aux informations. Cette étude a pu décrire la structure et les politiques sous lesquelles les cours et tribunaux sont assujettis, le mandat et les fonctions de la cour; ainsi que les fonctions spécifiques liées à l'objectif de cette recherche. De plus, l'étude révéla aussi le niveau d'implication de la gestion archivistique avec les quatre sujets de la recherche, les forces, les faiblesses et les lacunes liées aux quatre domaines de la recherche, et enfin les perspectives de stockage à long-terme des documents électroniques une fois l'informatisation et la numérisation du système sont mises en place. C'est ainsi qu'au cours des réunions qui ont eu lieu à Eldoret et à Arusha qu'on a pu examiner les conclusions. Au cours des échanges, on a élaboré un projet de stratégies, qui a été préparé par l'équipe de recherche pour résoudre les problèmes identifiés.

2. Vue d'ensemble des archives de la Cour suprême du Burundi et l'état de leur conservation

2.1. Photos prises pour les archives courantes du Bureau du greffe



Figure 1. Photo prise pour les archives courantes (Bureau du greffe).



Figure 2. Dossiers courants au Bureau du greffe.



Figure 3. Dossiers en cours d'instruction (Bureau des commis greffiers).



Figure 4. Au-dessus de cette armoire, ce sont les registres d'enrôlement des dossiers en situation critique.

2.2. Photos prises pour les archives définitives dites à classer (Cave de l'immeuble abritant la Cour suprême du Burundi)



Figure 5. Dossiers d'archives endommagés par les eaux de pluie passant par les trous et les fenêtres d'aération de la cave de l'immeuble abritant la Cour suprême du Burundi.



Figures 6 et 7. Les restes de dossiers endommagés par les pluies diluviennes passant à travers les fenêtres d'aération de la cave.



Figure 8. Vieux dossiers d'affaires récemment rangés sur les étagères octroyées par la Coopération technique belge (Appui institutionnel du Ministère de la Justice).



Figures 9 et 10. Cas des archives se trouvant au rez-de-chaussée du Bloc administratif de la Direction des affaires juridiques et du contentieux (Ministère de la Justice et garde des Sceaux).



Figure 11. Dossiers de la Cour suprême tombés par terre depuis plus de dix ans.



Figure 12. Dossiers d'affaires par terre (écroulement d'étagères de rangement des dossiers depuis plus de 10 ans).



Figures 13, 14 et 15. Dossiers de la Cour suprême abandonnés au rez-de-chaussée du Bloc administratif de la Direction des affaires juridiques et du contentieux.

3. Objectifs de l'étude

3.1. Objectif global

Les tribunaux traitent de grandes quantités de documents confidentiels, qui ont un grand impact sur la vie des citoyens et citoyennes, surtout en cette période postconflits où l'on remarque un accroissement spectaculaire des procès datant des époques les plus reculées et les plus saillantes de l'histoire tragique du Burundi (1962, 1965, 1972, 1988, 1993, etc.). L'objet de cette étude est d'acquérir une compréhension globale du fonctionnement de la Cour suprême du Burundi, d'abord en tant qu'organe suprême de la magistrature burundaise qui gère des dossiers sensibles et surtout en ce qui est de la gestion des dossiers d'affaires. De même, l'étude comprend aussi une enquête sur la gestion électronique des documents au sein de la Cour suprême du Burundi en vue de recommander un cadre qui peut être utilisé pour gérer les documents électroniques. Enfin, l'objectif de cette étude de cas judiciaire est de comprendre dans quelle mesure le domaine judiciaire s'occupe de la gestion de dossiers dans le milieu électronique.

3.2. Objectif spécifique

L'objectif spécifique de cette évaluation est de déterminer:

1. Le référentiel législatif et politique dans le cadre duquel les tribunaux fonctionnent;
2. Le mandat de ces tribunaux et les fonctions nécessitées par ce mandat;
3. Les règles, régulations et procédures qui guident l'exécution des processus et des fonctions des tribunaux;
4. Les fonctions spécifiques et les processus de tribunal qui font l'objet de l'étude de cas;
5. La mesure dans laquelle les quatre sujets du projet—la gestion de dossiers, les TIC, l'e-gouvernement et la liberté d'information—soutiennent les fonctions et processus individuels des tribunaux;
6. Les défis dans chacun des quatre domaines principaux concernant les fonctions et processus des tribunaux;
7. Les avantages offerts par chacun des quatre sujets du projet dans le cadre de leur application aux fonctions et aux processus des tribunaux;
8. Identifier qui est responsable de la gestion des dossiers au sein de la structure et quelle priorité donne-t-on à la gestion de dossiers;
9. La mesure dans laquelle la direction apporte son soutien aux bonnes pratiques dans le domaine de la gestion de dossiers;
10. Identifier quelles sont les capacités du personnel dans les tribunaux concernant spécifiquement
a) les effectifs dans le domaine de la gestion de dossiers, b) l'ancienneté, c) l'éducation, d) la formation, e) l'expérience;
11. Cibler quelles initiatives d'information mène-t-on dans les tribunaux, y compris les projets de numérisation;
12. La mesure dans laquelle le personnel de gestion de dossiers est impliqué dans la conception et la mise en œuvre des applications TIC dans les tribunaux;
13. La mesure dans laquelle les obligations de gestion de dossiers sont intégrées à une éventuelle informatisation des fonctions des tribunaux; et

14. Identifier quels dispositifs met-on en place pour la conservation et préservation des dossiers électroniques au long-terme.

4. Situation générale des archives judiciaires de la Cour suprême du Burundi

Les dossiers constituent à la fois des outils et des preuves sans lesquels les procès ne peuvent être plaidés et exécutés. Les dossiers d'affaires doivent être tenus avec beaucoup de soin, sinon leur disparition ou leur endommagement peut causer préjudice à l'une des parties au procès et à l'État avec des dédommagements. Ce qui est inquiétant est qu'en tant qu'organe supérieur de l'appareil judiciaire, la Cour suprême du Burundi ne dispose pas en son sein de service d'archives. Les dossiers en cours d'instruction sont tenus par les greffiers et sont conservés dans leur bureau à cause de leur utilité courante. Les dossiers clôturés, quant à eux, sont transférés dans la salle dite « Archives » où ils sont entassés et laissés à eux seuls sans aucun traitement.

Le manque de service d'archives s'accompagne aussi d'un manque de personnel qualifié. Ce problème est d'autant plus important que ces dossiers se détériorent de jour en jour en vue de tous ceux qui étaient censés les protéger. À côté de cela, aucun versement auprès des archives nationales n'a jamais été effectué, alors que le décret portant création du dépôt légal des archives nationales stipule que toutes les archives définitives de toutes les administrations, tant publiques que privées, devraient être transférées et conservées au sein de ce dépôt.

Aucun manuel de procédures de gestion des dossiers juridiques n'est disponible. Il n'existe pas de système de gestion de dossiers électroniques. Il n'existe pas de système informatique de gestion des dossiers. Il n'existe aucun responsable de la mise en œuvre de nouveaux systèmes de dossiers et d'informations. Il n'y a pas de fonction de gestionnaires de dossiers. Il n'existe pas de norme de métadonnées utilisée par le système informatique. Il n'existe pas de norme d'interopérabilité gouvernementale.

Quant à la recherche des dossiers, on peut trouver les dossiers ou informations contenues dans les archives courantes et semi-courantes (par ex. : les dossiers en cours d'instruction et ceux nouvellement transférés aux archives). Cependant, cela demande beaucoup de temps pour chercher là où ils se trouvent en ce sens que les documents transférés sont entassés pêle-mêle dans une salle abandonnée à elle seule appelée communément « archives » où seuls les plantons peuvent y accéder; elles sont poussiéreuses, couvertes de toile d'araignée et de boue causée par l'inondation (voir Figures 5, 6 et 7). Contrairement aux archives définitives (très anciennes), retrouver ces dossiers est quasiment impossible du fait que ces documents sont entassés et forment de montagnes de dossiers presque abandonnés.

Néanmoins, comme on a pu le constater à travers de cette étude, les archives de la Cour suprême du Burundi se trouvent dans une situation alarmante. Dans tous les endroits où nous avons pu passer, l'état des documents produits ou reçus est inquiétant, car ils font souvent face à de multiples problèmes liés : au manque d'équipements adéquats; au manque de personnel qualifié; à l'impossibilité de repérer l'information sensée existante; à la difficulté de maîtriser l'inflation documentaire; à l'absence d'une législation archivistique; à l'exiguïté des locaux; etc.

4.1. État de leur conservation

Les archives de la Cour suprême sont très mal conservées. Les documents sortis de l'usage courant sont entassés d'abord dans les armoires puis, quelques mois ou quelques années après, par terre dans les bureaux, les couloirs, et les caves. Parfois, ils sont placés dans des endroits impropres à la merci de tous

les agents destructeurs tels les termites, la chaleur, l'humidité, la lumière, l'obscurité, l'inondation, etc. Plus graves encore, les agents de la Cour suprême se débarrassent allègrement de ces « vieux papiers », car, pour eux, ils ne sont plus utiles. Alors que ce sont des documents importants et très précieux qui gisent dans les bureaux parfois exigus, inappropriés et vétustes.

Toutefois, étant donné qu'il n'y a pas de politique claire de conservation, depuis la création jusqu'à la destruction du dossier, ces derniers ne restent pas pour longtemps accessibles et utilisables. La plupart des dossiers se trouvant dans cette salle sont régulièrement détruits par la faune (rongeurs, insectes, etc.), des inondations des eaux de pluie passant par les trous d'aération situées à même le sol, obscurité, et bien d'autres intempéries (cfr Photos prises dans la cave dite « Archives »).

Pour les dossiers en cours d'instruction, il arrive très rarement que des pièces de dossiers soient perdues du fait des erreurs de rangement dans les fardes-chemises respectives, du fait des agents irresponsables qui s'adonnent le droit d'enlever des pièces dans un dossier soit pour les faire disparaître ou les remplacer par d'autres et par une destruction minutieusement préparée par les justiciables coupables de certains dossiers criminels (par ex. ce fût déjà le cas, où un justiciable coupable a causé un incendie à la Cour dans le but de détruire son dossier; heureusement que la situation a été vite maîtrisée).

Les magistrats instructeurs de dossiers ou leurs supérieurs hiérarchiques peuvent librement accéder à leurs dossiers, sauf aux dossiers sensibles, par exemple les dossiers de hauts dignitaires ou tout autre dossier pouvant porter atteinte à la Sûreté nationale de l'État. Ces derniers sont conservés dans les tiroirs fermés à quatre tours du Bureau du Président de la Cour suprême, qui seul en est responsable. Les dossiers judiciaires sont des dossiers sensibles dans la vie des peuples et des nations. Il existe néanmoins des inquiétudes sur leur conservation et leur élimination, car des dossiers mal conservés peuvent disparaître ou être détruits alors qu'ils détiennent toujours des informations pouvant servir de preuve tangible dans l'exécution des jugements rendus.

4.2. État de leur classement

L'état de classement n'y est évidemment pas meilleur. Que ce soit dans les bureaux du greffe et dans la cave de l'immeuble abritant le Parquet général de la République et la Cour suprême, il n'existe ni de cadre de classement, ni d'instrument de recherche afin de faciliter la communication. De même, il n'existe non plus ni de système de gestion papier organisé, ni de système électronique. Cependant, dans le but de faciliter un repérage rapide des dossiers courants, les greffiers se débrouillent en utilisant un système de classement souple et facile.

4.3. État de l'équipement

Les services du greffe de la Cour suprême du Burundi fonctionnent toujours de manière archaïque, sans matériel moderne de conservation et de communication. Aucun projet d'informatisation n'est à ce jour rendu possible dans ces différents services. Ce qui ressort de cette étude de cas est que la Cour suprême du Burundi est paralysée, elle ne peut accommoder aucun versement nouveau et ne peuvent pas matériellement aussi accomplir sa mission primordiale de collecte, de traitement, de conservation et de communication des fonds qui sont à sa disposition. La situation actuelle est vraiment décourageante.

Les greffiers sont amenés à entasser auprès d'eux les documents qui sont à leur charge, immobilisant ainsi des surfaces et des équipements de bureau. Comme ces documents gisent le plus souvent sur le sol faute d'équipement suffisant et adéquat, il est fréquent qu'on passe des heures et des jours à chercher sans succès un dossier indispensable.

4.4. État des locaux

Un autre véritable problème des archives de la Cour suprême du Burundi est celui du manque des locaux spacieux de conservation qui fait entorse à tout accroissement du fonds d'archives. Certaines archives de la Cour Suprême du Burundi, par exemple, sont installées dans de vieux bâtiments coloniaux qu'elles partagent avec les documents du Centre d'études et de documentation juridiques, tandis que d'autres sont entassées pêle-mêle dans une salle dite « Archives » de l'immeuble abritant le Parquet général de la République et la Cour Suprême. De même, les bureaux administratifs des greffiers sont exigus de sorte qu'ils ne peuvent pas accueillir d'autres documents courants.

4.5. État du personnel

La Cour suprême du Burundi ne comporte pas de spécialistes de dossiers, car ce sont les greffiers, dont leur profil est de niveau de diplôme des humanités complètes (BAC), qui sont chargés du classement et du rangement des dossiers sur les rayons. Autrement dit, il n'existe pas des fonctionnaires spécifiques auxquels la gestion des dossiers est assignée. Mais la gestion courante des dossiers est assurée par les greffiers sous la supervision du greffier en chef. Il est à noter que certains dossiers de par leur caractère sensible à la Sûreté nationale sont gérés par le Président de la Cour suprême (par ex. : dossiers Hussein Radjabul; dossier de l'assassinat de feu Président Ndayae Melchior; dossier des Puchistes de 1993, etc.) S'agissant du personnel formé, la Cour suprême ne dispose d'aucun spécialiste formé en sciences de l'information documentaire, en bibliothéconomie ou du moins en archivistique.

5. Approche de solutions (Stratégies à mettre en œuvre)

La situation que nous venons de constater exige un certain nombre de mesures immédiates de déblocage et, en même temps, l'établissement d'un plan permettant de résoudre définitivement le problème. Néanmoins, à force que la gestion des dossiers avance en relation avec les technologies de l'information et de la communication, l'e-gouvernement et l'accès aux informations, *l'International Records Management Trust* en collaboration avec le Secrétariat de l'*East African Community* et l'*East and Southern African Management Institute* (ESAMI) ont organisé une réunion internationale à Arusha le 1 au 2 septembre 2011.

Cette réunion a rassemblé divers participants clés venus des cinq États membres de l'EAC, ainsi que des intervenants internationaux en provenance de Londres, du Canada, de Norvège et des États-Unis. Les secrétaires permanents chargés de la réforme du secteur public, des TIC et des archives nationales, les archivistes nationaux venus des cinq États membres de l'EAC, ainsi que les responsables de l'EAC étaient aussi parmi les invités. Au cours de cette réunion, on a examiné les conclusions et on a proposé au cours des échanges un projet de stratégies, préparé par l'équipe de recherche, pour résoudre les problèmes identifiés.

5.1. Texte législatif (loi archivistique)

Cette loi n'existe pas encore au Burundi. Seul le Décret no 100/49 du 14 mars 1979 portant sur la création du dépôt légal des archives de la République du Burundi. Toutefois, théoriquement et selon le décret ci-haut, ce dépôt d'archives est censé recevoir les versements d'archives non courantes dites « à classer »; et veiller aussi à la conservation entière des archives publiques courantes et semi-courantes dans les divers services et administrations producteurs (dépôts provinciaux).

Il n'y a pas une autre législation qui ne concerne l'obligation d'archivage et de la conservation et la protection des archives nationales, ni dans le secteur judiciaire, ni dans d'autres secteurs. Étant donné que ce texte juridique est devenu confus, caduc, lacunaire, obsolète, incomplet au regard des autres textes régissant les archives à l'étranger, il faudrait plaider dans l'immédiat en faveur de sa mise à jour afin qu'une loi archivistique plus efficace et opérationnelle soit une réalité au Burundi. Dans ce cas, le législateur devrait s'inspirer de la loi étrangère afin d'adopter un modèle d'organisation administrative plus efficace. La mise en place d'une telle législation porterait sur : la prescription du versement des archives aux Archives nationales; la communication des archives; la création d'un conseil supérieur des archives; etc.

5.2. Évaluation

Il s'agit de l'évaluation des archives de l'administration publique pour connaître la quantité, la qualité des archives produites et leurs conditions de conservation. La mission d'évaluation consisterait à l'état des lieux de l'archivage dans l'administration publique à travers tout le territoire national. De plus, au regard de la situation actuelle, cette évaluation consisterait à sortir un rapport global dans lequel seront proposées des solutions pour la sauvegarde du patrimoine archivistique au Burundi.

5.3. Formation du personnel

La plupart des archivistes Burundais sont formés à l'Institut Supérieur de Commerce (ISCO) de l'Université du Burundi, option : bibliothéconomie. C'est une institution universitaire spécialisée dans la formation des professionnels de l'information incluant en plus des archivistes, des bibliothécaires et des documentalistes. La formation donnée à l'ISCO se compose d'un seul cycle universitaire d'une durée de deux ans (BAC+2). Cette formation est sanctionnée par un Diplôme d'études supérieures en bibliothéconomie. Il convient d'avoir à l'esprit que le nombre d'archivistes au Burundi est, à l'heure actuelle, très limité. Comme l'État n'offre pas de bourses de formation dans ce domaine, on pourrait procéder à des formations par des stages de courte durée et des formations continues au niveau local, régional et à l'étranger.

5.4. Coopération internationale

La coopération internationale dans ce domaine entre le Burundi et ses partenaires privilégiés comme le Conseil international des archives (ICA), CENARBICA, etc., n'est pas au beau fixe à cause du non-paiement des cotisations. De même, dans les protocoles d'accord signé dans le cadre de la coopération bilatérale et multilatérale, le domaine des archives ne fait pas partie dans les priorités du gouvernement burundais. De plus, d'après cette étude, les NTIC ont timidement pénétré dans certains services publics et privés au Burundi, y compris la Cour suprême, sans oublier la modernisation dans la gestion des archives burundaises. Cela est dû aux coûts exorbitants des logiciels commerciaux, à la faible vulgarisation des logiciels gratuits, au manque ou à la rareté du personnel spécialisé, au manque de matériel informatique et au manque de volonté des décideurs qui pensent que la gestion des archives ne relève pas des priorités du moment.

Néanmoins, pour relever donc ces défis, la coopération internationale pourrait faire des technologies de l'information un cheval de bataille en vue de faire des archives un outil de bonne gouvernance, de transparence administrative, d'efficacité administrative, et d'outils de prises de décisions

stratégiques pour l'amélioration de la productivité, la capitalisation des connaissances, la promotion du travail collaboratif.

6. Recommandations issues d'une réunion d'Arusha

À l'issue de cette réunion internationale qui s'est tenue à Arusha, on a pu identifier les fonctions spécifiques liées à l'objectif de cette recherche, le niveau d'implication de la gestion archivistique avec les quatre sujets de la recherche, les forces, faiblesses et les lacunes liées aux quatre domaines de la recherche et enfin, les perspectives de stockage à long-terme des documents électroniques une fois l'informatisation et la numérisation du système mises en place.

Spécifiquement, les recommandations formulées portaient sur:

- Les possibilités de partenariats entre les organismes responsables de priorités stratégiques du gouvernement, comme les TIC, l'e-gouvernement et l'accès libre aux informations;
- Les conseils devraient élaborer et développer un cadre de gestion pour les dossiers liés aux initiatives TIC, à l'e-gouvernement et à l'accès libre aux informations;
- Les stratégies liées au renforcement des capacités à l'endroit des ressources humaines devraient être développées au sein des organisations responsables de la gestion des documents, des TICs, d'e-gouvernement et d'accès libre aux informations;
- La faisabilité d'établir un centre d'excellence pour la gestion des dossiers à l'ESAMI devrait être explorée et des mesures devraient être prises;
- Un agenda pour le développement et l'adoption de normes, pratiques, procédures, systèmes et outils doivent être élaborés sur la base de la création à court, moyen et long-terme;
- La sensibilisation à tous les niveaux des stratégies et des outils à développer par les pays membres de l'EAC, ainsi que dans d'autres pays à travers le monde devrait, est d'une impérieuse nécessité;
- Un modèle de composants d'une politique pour la gestion des dossiers devrait être développé et pourrait être utilisé dans toute la région;
- Aligner l'archivage avec les TIC, l'e-gouvernement et l'accès aux informations dans tous les pays de l'EAC;
- Des normes standards de numérisation internationalement approuvées qui abordent le statut et la gestion des documents électroniques et de documents papier numérisés devraient être évaluées et adaptées au besoin pour assurer l'intégrité des enregistrements associés à ces initiatives;
- Un plan de préservation numérique devrait être développé pour assurer la préservation de ces documents électroniques qui doivent être conservés à long-terme;
- Renforcement de la gestion des documents, des TIC et de l'e-gouvernement dans la région de l'EAC afin de changer la façon dont l'administration publique travaille;
- Renforcement des capacités en matière de gestion des dossiers, des TIC, et de l'e-gouvernement;

- Amélioration des services offerts aux citoyens via la gestion efficace des documents et l'accès aux informations grâce à l'usage des TIC et de l'e-gouvernement;
- Le développement des infrastructures modernes pour pouvoir permettre aux citoyens un accès facile et aisé à l'information;
- L'adoption et l'harmonisation d'une législation sur la gestion des documents, les TIC, l'e-gouvernement et l'accès aux informations;
- La gestion efficace des documents électroniques pour garantir la fiabilité de l'information, car dans la plupart du temps, ils sont fragiles;
- Renforcement du partenariat entre les parties prenantes comme les archives nationales et les TIC;
- Les stratégies de gestion des documents, des TICS, de l'e-gouvernement, et de l'accès aux informations devraient refléter les questions nationales et régionales en identifiant les buts à court terme qu'on peut réaliser le plus vite possible;
- Une mise en place effective d'un comité technique de suivi au niveau de chaque pays de l'EAC est d'une impérieuse nécessité; et
- Création de ce comité technique au niveau national pour des rencontres régulières.

6. Conclusion

Pour conclure, la Cour suprême du Burundi est la plus haute juridiction du pays. De ce fait, elle reçoit en cassation des dossiers n'ayant pas pu obtenir gain de cause au niveau des juridictions inférieures, ainsi que les dossiers spéciaux de hauts dignitaires politiques. Une attention particulière dans la gestion de ses dossiers serait souhaitable en vue de protéger toute une masse de dossiers sensibles reçus régulièrement au sein de cette cour.

Nos impressions sur la recherche et les quelques interviews recueillies auprès du personnel et des justiciables rencontrés à la Cour suprême nous permettent d'émettre quelques recommandations:

- Dans l'immédiat, l'inventaire et la collecte des dossiers en péril dans la cave de l'immeuble abritant la Cour suprême et dans le rez-de-chaussée du Bloc administratif de la Direction des affaires juridiques et du Contentieux s'avère d'une nécessité impérieuse;
- La recherche d'une salle de conservation des dossiers d'archives de la Cour suprême pour répondre à l'exiguïté des bureaux où sont classés les dossiers courants, semi-courants et définitifs;
- L'informatisation et la numérisation du système d'information de la Cour suprême sont nécessaires pour faciliter la recherche d'informations et la pérennité dans la conservation des dossiers; et
- La création d'un service d'archives avec un personnel qualifié peut aider dans la promotion des archives de la Cour suprême.

Ce qui est étonnant, le Burundi va fêter le 1^{er} juillet 2012, le cinquantenaire d'indépendance sans aucune amélioration en matière de préservation de la mémoire collective nationale. De toutes ces recommandations, espérons-le, pourront marquer les jalons sur les orientations à donner dans la gestion

d'archivage, les technologies de l'information et de la communication, l'e-gouvernement et l'accès aux informations pour une meilleure connaissance des dossiers de la Cour suprême. Ainsi, nous nous permettons de clore notre présentation en espérant une nouvelle page dans le développement de la Cour suprême et dans tous les services tant publics et privés du Burundi.

Digitizing A Survivor's Identity

The Past, Present, and Future of the Kuwait National Museum Archives

Farah Al-Sabah

Kuwait National Museum, farahalsabah@gmail.com

Abstract

Kuwait National Museum was established in 1957. Since then, it has accumulated its massive archives, which have survived a devastating war. The decision in 2009 to digitize the ID cards, photographs, manuscripts, and other collections in the archives has resulted in the implementation of the ADLIB museum database system. This manuscript will recall the history of the archives at the Kuwait National Museum, how they were affected during the 1990 Iraq invasion, and how they were united and digitized. A special concentration will be given to the implementation of the ADLIB software program, explaining our process and challenges.

Author

Farah Al-Sabah has been a Kuwait National Museum employee since 2006. Although primarily working as an archaeological conservator, Ms. Al-Sabah was asked in 2009 to join an inventory commission of the Museum. After presenting the findings and offering solutions, Ms. Al-Sabah was asked to head a newly created team which would be digitizing the archives of the Kuwait National Museum. Farah Al-Sabah also participated in the ATHAR and SOIMA programs of ICCROM, as well as participating in other international UNESCO conferences.

1. Introduction

Everyone, everywhere, has been asked to identify themselves. ‘What’s your name?’ ‘Where are you from?’ ‘What do you do?’ These are questions we have all been asked, as well as asked other people. It is how we identify how this person fits in our lives. It is how we judge what purpose he/she serves. It is how we decide if he/she is important. The very same questions of identity are used on a country’s cultural heritage. Museums all over the world showcase their importance by displaying their unique contributions to mankind. Destroy these objects of pride and you have a nation robbed of its history. Lose the archives of a museum and you can make it seem like a nation never existed. The archives of the Kuwait National Museum contain identification cards, photographs, audio-visual recordings, and historical documents. The archives are a tangible link to our past, holding detailed information about an otherwise obscure archaeological artefact, or a first-person account of what life was like 100 years ago. The bulk of the archives have survived relocation, war, neglect, and have finally been granted digitalization, the closest thing to immortality.

This paper will discuss the history of the archives at the Kuwait National Museum, with a special concentration on the identification cards and their digitization. Thematically, the history of the archives will be separated into pre-war, war, and post-war, with a brief explanation of what international laws and conventions Kuwait is obliged by. The focus of the paper will be on the post-war unification of the archives, the eventual digitization process, the use of the ADLIB database program, and the various challenges that were presented.

2. Kuwait National Museum

The Kuwait National Museum was established in 1957. Kuwait's Amir, Sheikh Ahmed al Jaber al Sabah, granted his residence to the State, on the enviable shoreline of the city. In 1976, some artefacts were moved to a traditional Kuwaiti house near the Museum, so that the building would be constructed to serve the purpose of what it has become. When the Kuwait National Museum officially opened in 1983, archaeological and traditional artefacts were proudly displayed. It was during this time that the Museum took its responsibilities of cataloging the archives seriously, creating a responsible department.

The Kuwait National Museum is comprised of several sections, including the Archaeological Museum, the Heritage Museum, the Dar al Athar al Islamiya (House of Islamic Antiquities) Museum, and the Planetarium. Save for the semi-private Islamic Museum and the Planetarium (they are the responsibilities of differing departments), the artefacts in the Archaeological and Heritage museums all have identification cards that are archived.

The Heritage Museum contains a life-sized 'old town Kuwait' that visitors can walk around in. In the 'souk' part of the 'old town', traditional, often authentic, materials are used to convey a sense of what life would have been like. Items including clothing, tools, boxes, woodwork, swords, utensils, and other necessities housed in the Heritage Museum all have identification cards in the archives. Cataloging these often mundane items has proved rewarding, as the younger employees reading the ID cards now are reminded of the forgotten colloquial Kuwaiti words for items no longer used.

The Archaeological Museum, while less crowded with artefacts than the Heritage Museum, houses our most prized possessions. Rare and unique archaeological finds are proudly displayed, with information plaques boasting Kuwait's rich history. The information plaques would often read 'Excavated by Danish Expedition in Failaka Island, 1952'. The Kuwait National Museum was established in 1957, and had it not been for the research papers and meticulous record keeping of the excavation teams, the artefacts would have lost part of their luster. This would not be the last time that foreign assistance would be needed in order to verify our history.

3. Kuwait and UNESCO

The State of Kuwait became an entity on its own on June 19, 1961. One of the laws predating the 1961 Constitution was the Princely Decree on the Antiquities Law n. 11 of 1960.¹ Laws such as these helped to shape how important the Government of Kuwait viewed antiquities and cultural property. In fact, Kuwait was keen to protect its cultural property and heritage on an international level, as it was an eager signee to the various United Nations laws and conventions in this regard.

By 1963, Kuwait joined the United Nations, becoming its 111th member. Since then, Kuwait has been a fruitful member of the international body, and has "upheld the UN's principle of constructive cooperation, based on peace, equality and justice."² Kuwait has contributed economically to over 100 state-members of the UN who have sought assistance through the Kuwait Fund for Economic

¹ Copy of the "Princely Decree on the Antiquities Law n.11, 1960" can be found on www.unesco.org/culture/natlaws.

² "Kuwait and the United Nations," <http://www.embassyofkuwait.ca/pages/International/Kuwait-UN.htm>.

Development. To date, over 12 billion dollars have been given in assistance of the various members of the United Nations.³

Kuwait was also interested in contributing to the cultural aspects of the United Nations, as it joined UNESCO in ratifying and accepting the various cultural laws and conventions. The first example of such laws was the accession of the 1954 “Convention for the Protection of Cultural Property in the Event of Armed Conflict” law, popularly known as the Hague Convention.⁴ The ‘Hague Convention’ was the UN’s response to the devastating destruction of cultural heritage during the Second World War. In it, the ‘Hague Convention’ “sought to ensure that cultural property, both movable and immovable, was safeguarded and respected as the common heritage of humankind.”⁵ Kuwait’s accession to this groundbreaking UN Convention, as well as the subsequent “Protocol to the Convention”, meant that Kuwait’s cultural property had to be respected, especially since Iraq, which invaded Kuwait in 1990, itself had ratified the same Convention.⁶

The “Convention on the Means of Prohibiting and Preventing the Illicit Import, Export, and Transfer of Ownership of Cultural Property” of 1970 was also signed by both Kuwait and Iraq.⁷ The Convention helped cement Iraq’s “obligation to return exported cultural property” from Kuwait when it took the National Museum’s archives (among other irreplaceable property) in 1990.⁸ This would prove ironic, as although Saddam Hussein’s forces invaded Kuwait and took its national treasures, it was Kuwait which helped hold Iraq’s stolen objects for safekeeping during the 2003 US-led invasion.⁹

Kuwait’s willing hand to help, and with its peaceful and friendly nature, led the UN’s Security Council to unanimously condemn Iraq’s unprovoked invasion of Kuwait in 1990.¹⁰ Had the UN not intervened, especially since Kuwait was a signee to the UN conventions that protected cultural property, the archives of the Kuwait National Museum, along with other treasures, might never have been recovered.

4. The 1990 Iraqi Invasion

On August 2, 1990, Kuwait suffered an unprovoked invasion by its neighbor Iraq.¹¹ The subsequent looting of the Kuwait National Museum “was one of the greatest art crimes of the twentieth century.”¹²

³ Ibid. It should be noted that the amount of assistance is double in percentage that which was agreed on internationally.

⁴ 06/06/1969 Accession by Kuwait, “Convention for the Protection of Cultural Property in the Event of Armed Conflict with Regulations for the Execution of the Convention,” The Hague, 14 May 1954, www.unesco.org/eri/la/conventions_by_country.asp?

⁵ Rene Teljeler, “Chapter 9: Preserving cultural heritage in times of conflict,” in *Preservation Management for Libraries, Archives, and Museums*, ed. G. E. Gorman and Sydney J. Shep (London: Facet Publishing, 2006), 133-165, www.culture-and-development.info/issues/conflict.htm.

⁶ 21/12/1967 Ratification by Iraq, “Convention for the Protection of Cultural Property in the Event of Armed Conflict with Regulations for the Execution of the Convention,” The Hague, 14 May 1954, www.unesco.org/eri/la/conventions_by_country.asp?

⁷ 22/06/1972 Acceptance by Kuwait and 12/02/1973 Acceptance by Iraq, “Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property,” Paris, 14 November 1970, www.unesco.org/eri/la/conventions_by_country.asp?

⁸ International Committee of the Red Cross, “Rule 41. Export and Return of Cultural Property in Occupied Territory,” www.icrc.org/customary-ihl/eng/print/v1_rul_rule41.

⁹ Teljeler, p. 10.

¹⁰ United Nations Security Council Resolution 660 (1990).

¹¹ Kuwait, while tiny in comparison to Iraq, had offered economic support to Iraq’s on-going war with Iran.

By September 1990, the stolen artefacts from the Kuwait National Museum were exhibited on Iraqi television as ‘war booty’.¹³ The archives at the Kuwait National Museum were also taken.

The cultural invasion of the Museum and its antiquities was not surprising, recounted my colleague Mrs. Nawal al Failakawi, as the Iraqi National Museum staff had been at the Museum “learning how we run things” only a week prior to the invasion.¹⁴ In fact, Mrs. al Failakawi’s story is corroborated by Dr. Donny George, the Iraqi Director of Relations at the Iraqi National Museum.¹⁵ Mrs. al Failakawi remembers how prior to the invasion, employees from the Iraqi National Museum had visited Kuwait’s National Museum in order to learn how to properly package antiquities, how to best showcase artefacts in their display cases, and how to organize the archives.

According to Dr. George, “we got the orders from the Ministry of Culture to go and insure the evacuation of the Kuwait Museum” and “I myself made a video film for the two museums, the Kuwait National Museum and Dar al Athar al Islamiya.”¹⁶ Once Kuwait was liberated and Iraq was forced by UN Security Council resolution 687 to return the stolen goods, Dr. George “handed the UN representative two volumes for over 25,000 items that we had.”¹⁷

Seven months after the brutal invasion, Kuwait was liberated on February 26th, 1991 by UN-backed coalition forces. UN Security Council resolution 687, adopted in April 1991, required Iraq “to facilitate the return of all Kuwaiti property seized.”¹⁸ In May 1991, the UN’s coordinator overseeing the returning of the objects “found virtually the entire KNM and DAI (Dar al Athar al Islamiya) collections in the Assyrian Hall of the Iraq Museum in Baghdad.”¹⁹ Almost a decade later, in 2000, the UN Secretary-General remarked that there were still plenty of items that have not been returned to Kuwait, and that “priority should be given to the return by Iraq of the Kuwaiti archives... and museum items.”²⁰

“The Iraqi pillage of Kuwait...was a complete and utter destruction of Kuwaiti cultural heritage.”²¹ The retreating Iraqi army deliberately set fire to three of the Kuwait National Museum’s buildings, ravaging any remaining collections. Then Director of the Museum Dr. Fahad al Wohaibi estimated that the cost of reconstruction would be \$6 million dollars.²² Twenty-two years later, there are still missing artefacts and treasures stolen from the Kuwait National Museum. Mrs. al Failakawi cannot be sure what the exact number is, as some of the identification cards corresponding to the objects are also missing.²³

¹² Stephanie Goldfarb, “Lessons in Looting,” *ARCA (Association for Research into Crimes Against Art) Blog*, July 28, 2009, www.art-crime.blogspot.com/2009/07/lessons-in-looting.html

¹³ Ibid.

¹⁴ 17/07/2012 Interview with Mrs. Nawal al Failakawi, a colleague and current head of the Exhibitions Department. She has been working at the Kuwait National Museum for 23 years, and was one of the employees responsible for creating a list of missing items.

¹⁵ Donny George, “The Truth About the Kuwait Antiquities,” *SAFE (Saving Antiquities for Everyone) Blog*, August 13, 2010, <http://www.savingantiquities.org/donny-george-the-truth-about-the-kuwait-antiquities/>.

¹⁶ Ibid.

¹⁷ Ibid.

¹⁸ UN Security Council Resolution 687 Section D Article 15.

¹⁹ Jonathan M. Bloom and Lark Ellen Gould, “Patient Restoration: The Kuwait National Museum,” *Saudi Aramco World Magazine* (Sept/Oct 2000): 18.

²⁰ International Committee of the Red Cross, “Rule 41.”

²¹ Goldfarb, “Lessons in Looting.”

²² Bloom and Gould, “Patient Restoration,” p. 20.

²³ In a 2010 newspaper interview, Mrs. al Failakawi estimated that there are 487 missing treasures from the Museum. James Calderwood, “Nearly 500 Kuwaiti Artefacts Remain Missing After War,” *The National*, June 30, 2010, <http://www.thenational.ae/news/world/middle-east/nearly-500-kuwaiti-artefacts-remain-missing-after-war#full>.

5. The Finance Ministry Wants Answers

In May 2009, the Finance Ministry of Kuwait formally requested the complete inventory of the administrative and cultural holdings at the National Council for Culture, Letters, and Arts. The Kuwait National Museum, under the N.C.C.A.L. was therefore obliged to form a technical committee to inventory and evaluate the movable assets at the Museum, something which has not been done on this scale before.

In order to carry out the required, a team of Museum employees was chosen to dedicate their time and efforts for the task.. Our committee's objectives, under the orders of the N.C.C.A.L., were threefold:

- To photograph the heritage and archaeological holdings.
- To classify the objects by chronology and by ages.
- To store and save all the data electronically.

It was immediately apparent that we would be unable to complete this government-sanctioned request unless we united the archives.

6. Uniting the Archives

Our committee was given complete access to the archives in order to identify what movable objects we had at the Kuwait National Museum. Every cultural object in the Museum had an identification card with a unique 'Kuwait Museum', or KM, number attached to it. The challenge was to unite the original ID cards and make sure that the cards, and the objects they represented, were all accounted for.²⁴

In order to prepare an authentic list of KM property, we had to cross-reference and triple-check decades worth of archives. I was the point-person to a team of three that ultimately united and cataloged the Kuwait National Museum's archives. To do so, we needed a dedicated space and some electronic equipment. We used computers on a network, a scanner, digital cameras, printers, and external hard drives to save our work on. We also used Microsoft Excel, Adobe Photoshop, Microsoft Picture Manager, and Microsoft Paint softwares to database our findings and clean-up the scanned pictures.

The archive department mentioned that they had about 8,000 Kuwait Museum identification cards. Our first step was to number an Excel sheet from 1-8000, and highlight the numbers with original identification cards. Once complete, the ID cards were then scanned and corrected with Photoshop. The scanned copies of the ID cards also included header information that included 'National Council for Culture, Arts, and Letters', as well as 'Archives of the Kuwait National Museum'. Finally, we linked pictures (macro, in-situ, conserved) found on CD's from the various departments and added them to the massive folder with the numbered KM sub-folders.

The tedious work finally paid off, as we were able to ultimately unite 7,705 ID cards found from the archives. 337 of those ID cards had no identifying pictures, and the Museum photographer was asked to photograph those objects. The numerous colour-coded cross-referencing also helped us pinpoint the twenty 'lost' ID cards, which were not referenced in any of the post-invasion archives' list.

The 7,705 ID cards were, as required by the Kuwait National Museum, its governing body the National Council for Culture, Arts, and Letters, and the Finance Ministry, then collected and presented in

²⁴ The ID cards, which contained research information, were often transferred to archaeological sites, such as Failaka Island, which housed year-round local and international archaeological teams.

a massive catalog. The catalog was organized both chronologically and by historical ages. In order to preserve the original Kuwait National Museum archives, these catalogs were then distributed to the archives department for daily use, to the archaeologists who needed them on-site, and to the Kuwait National Library as a back-up.

7. A Need for Change

The uniting of the Kuwait National Museum's archives proved to not only be a long and frustrating process, but the defining factor in realizing that we, as Kuwait's National Museum, needed to operate using best practices. The days of handwritten notes on original archives had to end. The lack of information associated with the creation of the ID card had to be noted (previously, no date of creation or author's name was signed). Also, the ease in which the ID cards went missing, or were taken for research was too often and too high-risk. The ID cards were not 'checked out', and various uses for legitimate reasons made it difficult to properly return the ID cards to their chronological place in the archives.

These challenges have hindered the authenticity and safe-keeping of the archives at the Kuwait National Museum. Therefore, upon completion of our uncharted task, we whole-heartedly argued for an internationally-approved computerized system for the registration of new materials.²⁵ After consulting with experts, we finalized our decision to present the internationally-lauded ADLIB program as the future of archiving at the Kuwait National Museum.

8. Implementation of ADLIB

ADLIB, an internationally accredited and integrated system for managing museums collections thankfully had a local distribution office in Kuwait.²⁶ When we had presented our achievements to the Deputy Secretary General of the National Council for Culture, Arts, and Letters, we thankfully found a supportive decision-maker who pushed for the digitization of the archives as much our team did. The authorized dealer for the ADLIB program had already worked on other culture-related government institutions under the National Council's umbrella, most notably the Kuwait National Library. As such, the company responsible for ADLIB had experience working with the National Council, and the National Council, in turn, had found a respectful and helpful partner.

ADLIB is essentially a software that helped us digitally archive and manage our museum collection in a professional and multi-faceted way. ADLIB was designed to follow the 'CIDOC Guidelines for Museum Object Information', which gave it an authoritative, standardized, and internationally-approved stamp of approval.²⁷ Most importantly, ADLIB allowed us to customize the program to our requirements, and use it simultaneously in Arabic and English, without jeopardizing the standards set forth by the international museum community.

The contract between the National Council for Culture, Arts, and Letters and the authorized dealer for ADLIB in Kuwait stressed the importance of training the Museum employees. The company trained us on

²⁵ Paper-based archives would be still be part of the archives, as we would print the archives and collect them in catalogs.

²⁶ Government institutions would prefer to give contracts to companies with an authorized dealer in Kuwait, in order to secure maintenance and technical help.

²⁷ CIDOC is a committee at the International Council of Museums, or ICOM. ICOM is a partner of UNESCO.

how to use ADLIB, and tweaked the program to our specific Museum needs. We were also lucky to benefit from the company's experience across the Arab world when we wanted to comply regionally on agreed Arabic terminology.²⁸ Even after the training provided and months of inputting, we still find ourselves surprised by how we can further utilize the ADLIB program to our specific Kuwait National Museum needs.

With ADLIB, we were able to input the information found in the paper archives, attach scans of the original pictures, include updated high-resolution pictures, video biographies of the items, and any other information that related to the items very easily. Before ADLIB, we would never be able to concretely state how many objects we had in the Museum that were from the Bronze Age, for example. With ADLIB, and since our existing archive was based on four categories (Heritage, Hellenistic, Islamic, and Bronze Age), we were able to transfer our existing archiving categories to a computerized database. Using these four categories, we would not only be able to exact the total number of Bronze Age items, but also to pinpoint the items that were found in a certain trench on an archaeological site.

The hierarchical structure of ADLIB also allowed the users to find, for instance, how many wooden traditional boats we had in the Heritage collection. Using the predetermined hierarchy also forced the inputters and the users to uniform the terminology that was previously recorded on the paper archives as 'wooden boat', 'traditional dhow', or 'old Kuwaiti boat'. By using the same terminology as an identifier, we would be able to not only find out how many of the traditional wooden dhows we had, but to classify them according to date produced, condition, size, and other important factors. This hierarchy and classification would also be invaluable to the conservation and exhibitions teams, as it helped them decide what to work on or showcase.

Having all of these easily-accessible various options and descriptions in ADLIB made it the go-to answer for any Museum related query. Suddenly, the possibility of asking ADLIB a question about a Museum object and having a plethora of answers made members of the various departments want to use the program. Security was key when it came to physically protecting ADLIB from floods, blackouts, and viruses. However, we became aware early on during the inputting stages that we had to protect the system from the Museum employees themselves. Well-wishing employees from the storage department wanting to find out the exact location of an object could have possibly changed or altered the entry without our knowing. Therefore, we had to assign usernames and passwords for the inputters and to the department heads, which allowed us to set editing control and view editing histories.

Security and control of the content in ADLIB was also important for the Museum, who would offer its archives to members of the various archaeological teams or researchers. We would be able to hide or edit sensitive information that we would not like to be public, including location information, pricing, and donor information. Researchers working on Failaka Island who would request access to the archives would now have the option to print the information on the object they wanted to study as well as any object that would be of interest to them, made possible by using the previously-mentioned hierarchy.

9. The Future of the Museum and ADLIB

The contents of the Kuwait National Museum can be made available online, as ADLIB has the option of publishing the catalog to the internet. However, this option has not been considered because the digitization of information from the archives has not been completed. Also, we are hesitant to introduce

²⁸ The glossary for Arabic terminology in the field of cultural property has been long-discussed, with Dr Hossam Mahdy of the ATHAR program of ICCROM compiling the most used and extensive list.

an internet connection to the computers on the ADLIB network, for fear of inadvertently affecting the computers and the software with malicious content or debilitating viruses.

A major selling point of choosing ADLIB as the digital future of the Kuwait National Museum was the capability of adding more than just the objects at the Museum. ADLIB could be used to catalog the printed and audiovisual items housed in the Museum library, as well as the other museums under the control of the National Museum.²⁹ The far-flung museums in the desert and the island (Red Palace Museum in Jahra and Failaka Museum on Failaka Island respectively) could use the ADLIB program. The heritage homes museums such as Dickson House and Bait al Bader could catalog their items and add descriptions of how the houses were used.

The museums honoring our sea-faring traditions and the Kuwaiti Martyrs would also be able to catalog their collections by using ADLIB. Everything from the special pearl-diving tools found at the Maritime Museum to the blood-covered clothes worn by the Kuwaiti martyrs could be added to ADLIB easily, thus preserving the history of how items were used and who wore them.

Beyond the direct administration of the Kuwait National Museum, the prospects of digitizing and cataloging the Modern Art Museum's various multi-media artwork would showcase the re-vitalized art scene. Artwork otherwise condemned to the storage rooms would now be able to find new life in the online digital world. In fact, the National Council's Modern Art Museum would be able to compete for the art interest prevalent in the Kuwaiti and Arab scene, once the works of famous authors are shared on a global level.

Globally, we plan to coordinate with the international embassies of Kuwait to house and safe-keep the hard disks that we have backed-up. By doing so, we ensure that any copies of the Kuwait National Museum's identification card database and archive are protected should the Museum be compromised by any natural or man-made means. However, there are legal and ethical ramifications that we have yet to discuss with the decision-makers.

It is with these plans for the future that would not only catapult the Museum's archives to the 21st century, but would benefit the National Council for Culture, Arts, and Letters and the Government of Kuwait.

10. Technological and Economic Challenges

While the National Council for Culture, Arts, and Letters was supportive of the database collection and ADLIB implementation, the process was not without challenges. Other than the challenges that my colleagues and I faced with having to learn the ADLIB program, there were also technological, economical, professional, and ethical challenges.

A complex program such as ADLIB requires that the machines used are capable of seamless integration, which posed a technical challenge. While the ADLIB contract included three computers and a server, we wanted to ensure that the hard work of inputting the database would be safe from an erased memory or a crashed server. We faced difficulty in convincing the decision-makers that extra UPS protection (Uninterruptible Power Supply) was necessary to protect the computers from crashing should there be a power outage. The server is automatically set to download and save the day's work at 3am

²⁹ The library at the Kuwait National Museum should not be mistook as the Kuwait National Library, which incidentally uses the library version of ADLIB.

every night, but should the power suddenly go out, we didn't want to risk the computer hard drive from crashing.

Another technological challenge we faced was making sure that the database was properly saved and secured. As previously mentioned, the server automatically backs up the database daily, but we were adamant in securing the database on an external hard drive, as we could not risk the chance of a burst water pipe or a fire breaking out. Saving the entire database on an external hard drive is a task that is completed on a monthly interval, and the external drive resides in a separate building. Another way of saving the database was to prevent viruses of getting in. In order to protect the database, USB devices are not allowed to connect to the computers or networks, and there is no internet connection on the network system.

We were lucky, in a sense, of initiating the database in 2009. Starting on the database in this technologically advanced age saved us from facing obsolescence associated with floppy disks and microfiche. The challenge is to integrate into the future technology seamlessly. The ADLIB database is built on the Microsoft Windows operating systems, and has continuous support and upgrade conversions by the company. In fact, one of the most important clauses in the contract is the maintenance clause, which guarantees that the technology will be up to date.

The cost of implementing ADLIB, which was in the tens of thousands of dollars, was postponed until the proper funds could be secured. It would take two years until the contract was signed, and the prices of the updated equipment kept rising. Few could understand why a software program connecting a few computers was this expensive, or why we needed these add-ons of UPS devices, external hard drives, and virus protection. However, we made sure to include in the total cost of the contract any future maintenance renewals, machine upgrades, and software updates. In this way, we presented the National Council for Culture, Arts, and Letters with one big bill that they would only have to pay once.

11. Professional and Ethical Challenges

The Kuwait National Museum, where the archives are held and are being digitized, has plenty of dedicated, hard-working, and professional employees. The challenge, though, was to find all-around computer savvy and detailed database inputters. The ID cards are bilingual, and the inputters would have to be familiar with English in order to type it on a keyboard. The Arabic text also proved challenging, as some of the words were misspelled or were based on a colloquial term rather than a classical Arabic word. Certain Kuwaiti words written on the ID cards would not be found on other ID cards containing the same Heritage items.³⁰ Instances such as these would over-populate the database terminology hierarchy with two words for the same item, one in classical Arabic and one in the local colloquialism. Only a Kuwaiti, and a traditional-speaker at that, would be able to correct and edit as they moved along.

The Museum employees, having been kept aware of the database implementation, were invited to submit their names for the 'database input team'. The volunteers were made aware of the shared goals, and were asked to set aside two hours of their day to inputting. After securing the amount of inputters needed, and training them, we had begun the process of digitizing the Museum archives. Problems with the employee volunteering scheme began to quickly rise. Maternity leaves had to be granted, supervisors needed their employees back, vacations were scheduled, and other basic rights had to be upheld. This

³⁰ 'Kuwaiti' is simply a colloquial 'language' or accent of classical Arabic.

proved difficult in maintaining a steady stream of employees willing to work for no extra pay and without compensation.

An ethical debate took place when a suggestion was made for hiring professional database inputters. A quick discussion nixed the idea, as we were uncomfortable with handing over the Kuwait National Museum's database to foreigners. This outcome proved wise, as most of the Heritage ID cards could only be understood and corrected by a Kuwaiti. The decision to utilize the Museum employees would also prove beneficial to the colleague who works solely on manuscripts and correspondences.

12. Historical Documents

There are estimated to be about 60,000 historical documents present at the Kuwait National Museum. There were more, according to a colleague who handles the documents says, but a post-war 1992 clean-up of the Museum swept up hundreds, if not thousands of documents. While the historical documents are stamped with 'Kuwait National Museum Collection' and given a document record number, the exact number of documents missing is unknown. Taking these yellowed historical documents and digitizing them was a priority for our team during the implementation of the ADLIB program.

The historical documents cataloged are categorized according to their general topics, such as 'Ruling Family', 'Sea-Faring Documents', and 'Merchant Documents'. Of the total historical documents, it is estimated that only a quarter of them are digitally scanned. Of these digitally scanned documents, the bulk are early 1900's correspondence letters between Kuwaiti merchants in India and their families in Kuwait. Some of the correspondence letters scanned discussed family matters, and these have been marked by the manuscripts manager as classified. After customizing the program for us, the authorized ADLIB company was able to 'lock' the documents marked as classified, thus protecting the author's privacy.

Luckily for the ADLIB database input team, importing the historical documents files into the program is seamless. The manuscripts team had thankfully already added information with each scan, including title, date, to, from, and condition. We used this information to automatically populate the entries of the document record database by importing a Microsoft Excel sheet into the ADLIB program. The picture scans of the documents are in a .jpeg format, and the ADLIB program imports them into the corresponding number documents record. By digitizing the historical documents and adding them to ADLIB, the Kuwait National Museum will be able to account for these invaluable insights to Kuwait's past.

13. Digitization By The Numbers

The work of digitizing the archives has in some part started prior to our 2009 creation of a task force, and that has helped us immensely in stream-lining the digitalization process. I will provide below a table of what we have achieved thus far.³¹

³¹ As of July 20, 2012.

United KNM ID Cards (printed and bound to serve as an accessible paper archive—categorized by chronological order and by historical age)	7705
KNM ID Cards Missing from Archives	20
KNM ID Cards Without Pictures	337
KNM ID Cards Scanned	1236
KNM ID Cards Information Added to ADLIB	5229
KNM ID Cards Picture (scan of the original paper ID card) Added to ADLIB	5229
Pictures of KNM Objects Added to ADLIB	2994
KNM Document Records Added to ADLIB	3600

14. Recommendations for Other Institutions

As discussed throughout this paper, the process of digitization and parallel database inputting of the archives was only seriously undertaken at the Museum since 2009, at the orders of the Finance Ministry. Previous attempts at digitization or database inputting, while very commendable, have never come to fruition because of a lack of administrative, technological, and economic support.

While the Finance Ministry was the trigger for our work, had it not been the unwavering support of the now Secretary General of the National Council for Culture, Arts, and Letters, our Director of the Kuwait National Museum, and all the employees, our objectives would have been failures. You must ensure that you have the very top decision makers believing in your grand ideas. Detailed monthly reports were written to the General Secretary and the Director with progress, updates, and challenges faced by our team. The constant reminder of the work you and your team are doing will have more departments interested. This is a team effort, involving people you might never even meet.

As the Museum is part of the grand National Council, we would often have to rely on and convince people we have never had contact with. The paperwork you present must be extremely clear and easily understood by people outside the cultural discipline. Any requests for economic assistance must be crystal clear, with detailed explanations and justifications as to what exactly you are asking for and why. The finance and legal departments, populated by employees who have never considered the fate of the Museum archives or have seen the passionate presentations arguing for a digitized (united and uniformed) future, must be spoken to in their ‘language’ (clear budgetary numbers, referencing N.C.C.A.L. laws, etc.).

While the decision in choosing ADLIB was straight-forward and easy for us to make, remember that not all institutions are built the same. Make a list of what you need (ie: a united archive), in what capacity (ie: high volume of records with multiple media each), and who can provide it on a long-term (ie: local authorized company with reputed history). Make sure to take your time in deciding on a program or company, and only after thoroughly inspecting and vetting their previous work and reputation.

The most important requirement for a program or a company is for it to follow international standards set forth by the responsible and accepted museum or cultural institutions (ie: UNESCO, for one). There is absolutely no use in implementing an untested or unfamiliar database. Also, connect with

museum professionals around the world who have implemented databases for their archives. When you are doing something unprecedented in your country, you should network.

15. What's Next?

While we are proud of our achievements thus far in digitizing the archives at the Kuwait National Museum, we recognize, as a task force, that our work is far from over. There will undoubtedly be challenges that we have not considered, irreplaceable employees who will leave, and technology that will have to be upgraded.

There are still archaeological teams working every spring and winter in Kuwait, unearthing past civilizations and inevitably adding to the ever-growing identification cards archives. Attached to the newly added KM numbers are the corresponding pictures of the objects, which will certainly be photographed repeatedly in detail by the excavating teams. All of these on-going expected future works will have to be incorporated to the paper archives (and later bound chronologically and by ages) and ultimately to the ADLIB database.

The historical documents team has only digitally scanned about a quarter of the 60,000 strong historical manuscripts. If we are to truly have a complete and authentic archive, and not only the 'interesting' correspondences, then we must include the rest.

If the projected timeline holds, the Kuwait National Museum's objects are set to occupy custom-built storage buildings by 2013. Moving the objects to their newly dedicated space will require that the storage department and the ADLIB input team must work together to correctly identify the location of the objects. As such, we are brainstorming and gathering information into the possibility of utilizing electronic hand-held scanning devices which would date and time stamp the location of every object leaving the storages.

Over 60% of the current 1.3 million Kuwaiti population of Kuwait is under the age of twenty-five.³² As a new generation enters the government sector marketplace, taking the place of the veterans, we are constantly reminded of how valuable the employees with 20+ years in the Museum are. The veteran employees at the Museum know its history from before the invasion, helped rebuild it after liberation, and have learned from their mistakes. Their insight and expertise has been invaluable in uniting the archives, and their input into the implementation of the ADLIB workflow has helped us save months of unnecessary tedious work.

16. Conclusion

The archives at the Kuwait National Museum unintentionally came into existence when the then Amir of Kuwait endowed his residence in 1957 and declared it a museum. This paper has described the turbulent history of the archives at the Kuwait National Museum. The archives, and the invaluable identification cards that they hold, have mostly been unchanged and untouched for a minimum of thirty years, and are bare of any updates or changes to the history of the object (save for a few penciled-in notes). The

³² 2012 UNDP Kuwait Youth Survey, <http://www.undp.org/content/kuwait/en/home/presscenter/articles/2012/05/17/national-youth-survey-conducted-by-undp-kuwait/>.

historical documents, some of them dating to the late 1800's, have survived being forever lost in the clean-up of the Museum after the 1991 liberation. But by finally uniting and digitizing the contents of the archives at the Kuwait National Museum, we hope to lay to rest any fears of future discord.

The story of the archives at the Kuwait National Museum is nothing short of a survivor's tale. The 8,000 + archaeological and heritage objects all have identification cards, describing the object's story. Losing the paper-based identification cards would be nothing short of a tragedy, a loss of identity. By digitizing them, the Museum has finally honored them. And while we must always remember that digitization and technology is not the be-all and end-all of our problems, it is certainly the best step we have taken in order to protect our archives and our history.

The Challenges of Manuscript Preservation in the Digital Age

Wayne W. Torborg, Theresa M. Vann and Columba Stewart

The Hill Museum & Manuscript Library, www.hmml.org

Abstract

The Hill Museum & Manuscript Library (HMML) partners with libraries around the world to digitize and catalog their manuscript collections, both to ensure archiving of contents and to facilitate access by researchers. HMML is a not-for-profit organization that provides these services to partner libraries at no cost. HMML trains and pays local workers in each location; owning libraries retain publication and commercial rights to images of their manuscripts; HMML can make copies available to researchers subject to those conditions. HMML maintains long-term relations with each partner library so that its digital archive remains current and always available. HMML is currently working in several countries in the Middle East, as well as in Malta, Ethiopia, and India with the generous support of the Arcadia Fund and other sponsors.

Authors

Wayne W. Torborg has been an imaging specialist for over 30 years. After receiving a degree in mass communications in 1984, he operated a commercial photography business. In 1997, he was recruited by a digital media services company to produce digital photography for advertising clients. In 2004, he became director of digital collections and imaging for the Hill Museum & Manuscript Library (HMML) at Saint John's University in Collegeville, Minnesota. There he developed the Library's best practices for its overseas manuscript digitization studios and the long-term preservation of its digital assets. In his work at HMML, Torborg also supervises overseas HMML's manuscript digitization projects, manages its websites and databases, and is responsible for the preservation of over 140 terabytes of digital information.

Theresa M. Vann received her Ph.D. in History from Fordham University in 1992. She accepted the position as the Curator of the Malta Study Center at the Hill Museum & Manuscript Library in 1995. In the next five years she created on-line finding aids for the Malta Study Center's collections, became involved in developing the Library's online manuscript catalogue, and closed down the microfilming project in Malta. She opened a digital studio in the National Archives of Malta in 2007, and supervises the Malta Study Center's digital studios in Malta and Rome. She teaches medieval history at Saint John's University.

Father Columba Stewart, OSB, has been the executive director of The Hill Museum & Manuscript Library (HMML) since 2003. A native of Texas, Father Columba received his A.B. in History and Literature from Harvard College, an M.A. in Religious Studies from Yale University, and his D. Phil. in Theology from the University of Oxford. He is a professor of Theology at the Saint John's School of Theology, and has published extensively on monastic topics at both popular and scholarly levels, including *Prayer and Community: the Benedictine Tradition* and *Cassian the Monk*.

1. Background

The Hill Museum & Manuscript Library was founded in 1965 as a sponsored program of Saint John's University and Abbey in Collegeville, MN. The Benedictine community, shocked by the widespread destruction of libraries and archives in Europe during World War II, decided to revive the monastic

tradition of preserving knowledge for future generations by copying medieval manuscripts. Unlike their medieval predecessors, however, the 20th-century monks of Saint John's would use the best modern technology to copy manuscripts, which in 1965 was the 35-mm microfilm camera. Initially, the Library's preservation mission was limited to monastic libraries, but the scope gradually broadened to include all manuscripts produced before the year 1600. HMML microfilmed the manuscript and archival collections of four national libraries (Austria, Portugal, Malta, and Sweden), and worked with numerous universities and dioceses to preserve their collections. Unlike other projects that photographed the "prettiest" manuscripts, the Library established the practice of microfilming each folio in every manuscript in the collection based upon the principle that the collections should be preserved in their entirety, for the use of present and future researchers. By 1999, when the Library ceased its microfilm operations, HMML field directors had microfilmed over 90,000 manuscripts from Austria, Germany, Spain, Portugal, Ethiopia, Malta, Sweden, Switzerland, South Africa, and some Italian and English libraries.

From its beginning, HMML's mission was to preserve endangered manuscripts and make them accessible to researchers, giving priority to collections at risk from warfare, neglect, and/or lack of resources. Sometimes this is easier said than done. HMML is not a commercial vendor, and it does not charge a fee to libraries for photographing and archiving their collections. This is a great benefit to librarians and archivists who find themselves in a situation where their government or institution has no budget whatsoever for preservation. But some collections can remain inaccessible because of the current political or religious climate in the country; for example, the Library began talks with the Armenian Patriarchate of Istanbul in the 1970s to photograph its manuscripts. Work did not begin until 2005, and even then the project was so politically sensitive that the Library did not announce it until it was completed in 2009. Most frustratingly, sometimes one individual can delay a project for decades because of fear, ignorance, or just plain contrariness. Fortunately, HMML is an institution, not one person. If diplomacy and discretion does not work, the Library can afford to wait until circumstances change.

In order to secure the trust of its partner institutions, it is vital that HMML recognizes and protects the ownership rights of its partner libraries. In a similar spirit, HMML freely shares its cataloguing information online with its partner libraries. HMML's on-going relationship with its partner libraries is reflected in its organizational structure, which has curators and cataloguers responsible for each of its main collections: Austria-Germany, Eastern Christian, Ethiopia, and Malta. Finally, the Benedictine value of stewardship encourages HMML to take the long view of manuscript preservation and usage. HMML staffers are aware that the original manuscript takes precedence over its reproduction, and that their work does not replace the original.

The Library's transition to digital imaging in 2003 required a complete reassessment of its procedures. Microfilm technology determined the Library's internal organization, policies, and even its physical environment. HMML's transition from microfilm to digital images required the staff of the library to rethink how it implemented its mission and policies. Traditionally, manuscript libraries have been inaccessible to anyone except a small circle of specialists. First, users had to learn how to read a manuscript. Second, users had to find a manuscript to work on. Medieval manuscripts have intrinsic value, and the cheapest way to protect them from thieves is to keep knowledge of them secret. Then, unlike cataloging a printed book, cataloging manuscripts requires considerable expertise beyond mere research. If a researcher found a manuscript in a printed catalog, he or she had to travel to the manuscript. There, the intrepid researcher had to present credentials and obey the rules of the manuscript keeper to gain access.

Microfilm is an analogue copy of the original, so it was simple to treat a microfilm as if it were a manuscript. One microfilm was the equivalent of one manuscript. A reel of microfilm was labeled as one unit. Duplication could be easily controlled, because the technology of duplication was beyond the means of most individuals. Access could also be controlled, much like the original. Researchers still had to travel to the microfilm depository to look at a manuscript. However, a particularly valuable or fragile manuscript could be preserved by limiting researchers to using the microfilm.

Microfilm was the archival standard for preservation photography, and its best practices were already established when HMML began its work. The Library sent its field director to the site with a microfilm camera, stand, and the undeveloped microfilm. The field director prepared a microfilm inventory card for each manuscript, which assigned each manuscript a microfilm project number, and contained some rudimentary catalog information. The inventory card was filmed with the manuscript, and then filed as a finding aid. HMML tracked the microfilm through its processing, payment, storage, and retrieval by its project number. The industry promised that the microfilm would survive for 400 years under optimum storage conditions.

The field directors worked on site, and the Library was limited in the number of projects it could undertake at any one time. Cost was a major factor, because microfilm projects had large upfront expenses. In 1970 the Library's director concluded that a four-month project filming 1,000 manuscripts in Malta would cost \$45,763, with \$25,000 allocated for photographic materials and film development. This included the heavy-duty microfilm camera and its monumental stand along with the purchase of black and white microfilm stock. Still, the Library raised money and by 1973 it had three microfilm cameras in the field: in Spain, Malta, and Ethiopia. The Spanish project moved the camera from one library to the next, ensuring the security of the manuscripts; in Malta and Ethiopia, however, the camera stayed in one place and the materials came to it.

Personal computers and the Internet changed how the Library served scholars. It put up its first web page in 1995, and began a project (funded by the Andrew W. Mellon foundation) to create electronic standards for cataloging medieval manuscripts. Its OPAC (electronic catalog), now called "Oliver" after HMML's first field director (who said that electronic library catalogs would never catch on) went up in 1999. The Library provided its electronic assets free of charge, and the majority of its users accessed its website and catalog from computers in Europe.

The Library still microfilmed manuscripts, however, in black and white bitonal microfilm. The staff weighed the advantages and disadvantages of changing to digital. Microfilm had a proven archival track record; the longevity of digital technology was unproven. In comparison, digital images were in colour, and were easier to access than the black and white microfilm. Also, digital photography studios were cheaper to set up and the photography itself was inexpensive, in comparison with the costs of film and development. But once the microfilm has been developed, put on a reel, and boxed up, it could be stored in a cool dry place indefinitely for a low cost. In comparison, processing a digital asset was initially extremely cheap, but one also had to consider the ongoing costs of maintaining electronic data and its migration to new platforms. Furthermore, the staff had to change their method of conceptualizing and describing manuscript analogs, because if one manuscript equaled one reel of microfilm, the same manuscript equaled 300 digital image files. Finding a way to assign these digital image files sequentially to the correct manuscript required a new way to organize the filming projects. Despite these issues, in 2003 the Library took a deep breath and switched over.

2. Transitioning from Analogue to Digital Photography: Methodology and Rationale

At the turn of the 21st century, libraries, cultural preservation institutions, and image repositories confronted a new reality in the way visual information was gathered and disseminated. For decades, analogue technologies such as traditional photography and microfilming had served the needs of institutions needing to store visual information and individuals needing to access it. This changed as digital technologies for the storage and distribution of information matured and became more affordable.

Much of the impetus for this change hinged on the rapid adoption of personal computer technology in the 1990's combined with the rise of the Internet as a means of communication and information access. Researchers wanted digital images that they could use on their computers or access via online resources.

At that time, the Hill Museum & Manuscript Library (HMML) had a 35-year history of preserving manuscript images on 35mm black and white microfilm. HMML's leaders understood that digital imaging would offer numerous advantages for manuscript preservation and access, but needed to adopt appropriate digitization strategies to achieve success.

Most of the manuscripts imaged by HMML are bound codices, which present a number of challenges in digitization. They are often delicate and require careful handling to avoid damage. Many cannot be pressed under glass (the usual practice in the microfilming days) without damaging the book's binding. The efficient digitization of bound books is a challenge for institutions wanting to digitize collections and inventors and manufacturers trying to devise the best methods for doing it.

Any methodology for digitization represents a compromise between quality (image resolution, colour space, bit depth, etc.), efficiency and cost. Efficiency and cost are of particular importance to HMML—its manuscript preservation goals involve operating multiple digitization sites employing local technicians. This work often takes place in remote locations and under challenging environmental conditions. What was needed was a system that:

1. Would produce full-colour digital images of sufficient resolution and quality.
2. Would stand up to heavy daily use without breaking down, and be easy to maintain.
3. Was able to provide excellent productivity.
4. Would be priced so that multiple digitizing sites could be operated simultaneously.
5. Was easily transported and assembled.

There are many innovative, high-quality systems for the digitization of bound books. Flatbed scanners exist that are designed primarily for scanning bound volumes, but they are somewhat delicate and require excessive handling of the often-fragile codex. Specialized overhead scanners for books are made by several companies, but they are often quite large, difficult to ship, and rely on proprietary technology and software. They are also quite expensive and difficult to service once installed in a remote location. The most sophisticated book imaging devices available are robotic scanners that can scan and turn pages in a bound book automatically, but they are prohibitively expensive and would likely damage the types of manuscript materials HMML works with.

In considering its options, HMML looked to a pragmatic imaging system of the past: the 35mm camera mounted on a copy stand. In 2003, the camera manufacturer Canon introduced a new digital single-lens-reflex (dSLR) camera called the 1Ds. It was the first of a line of digital cameras employing a "full frame" 35mm sensor. It was styled and constructed very much like its rugged 35mm predecessors and produced a digital image measuring 4064 x 2704 pixels in size. This is an image of sufficient size for high-quality, full-page magazine publication, and is quite a bit larger than needed for any sort of online

presentation. HMML decided that a modernized copy stand photography system using this new camera could be the basis for practical book digitization.

HMML's manuscript digitization system consists of a dSLR camera, copy stand, freestanding electronic flash lighting units, custom-made tools and book cradle, and a personal computer. This system has proven to be reliable, easy to use, and produces high-quality photographic images of manuscript pages.

Innovative, cost-saving thinking has gone into the design of the system. High-quality copy stands are quite expensive, so HMML purchases used stands at low prices and modifies them to make them more portable and versatile for book photography.

The HMML book cradle is another example of this. It is made of foamboard, fabric, gaffer tape and Velcro® strips. Inexpensive to construct, it can be configured in any number of ways to hold books for photography, folds up for shipping, and weighs almost nothing. Commercially-made accessories for book photography are almost non-existent, so HMML fashions its own tools inexpensively using acid-free matte-board, clothespins, hot glue, clamps and clear plastic.

For lighting the book pages, the HMML system uses a pair of freestanding studio electronic flash units which can be obtained at a cost of \$300 each. Flash illumination offers a number of advantages over continuous lighting systems:

1. Unlike traditional tungsten or tungsten-halogen floodlights, flash units produce almost no heat. Heat generated by traditional "hot lights" can make a small enclosed studio uncomfortably hot while also drying the air, which can be harmful to parchment book pages.
2. The flash units emit "full spectrum" light and have very consistent power output from image to image. They tend to hold this consistency despite the electrical power fluctuations that are quite common in the places HMML works.
3. The brightness of the flash and the fact that a relatively high camera shutter speed and small aperture can be used means that ambient light is generally irrelevant to the photography. In practical terms, it means that a studio can be set up in almost any sort of space that's out of direct sunlight and that a special room or environment for the system does not have to be created at the studio site.
4. Using flash lighting virtually eliminates the possibility of producing blurred pictures due to camera vibration or movement. The effective duration of the burst of light produced by the flash is in the range of $1/3000^{\text{th}}$ of a second. At this speed any sort of movement of the camera or subject is "frozen."

HMML's system is adaptable. If a standard photographic tripod is added, the studio system becomes even more versatile, able to photograph large wall-mounted items, materials held in special fixtures, or small three-dimensional objects.

At one point, HMML had 20 of these manuscript digitization systems in concurrent use. This was possible in large part because of the relatively low cost of the system. With a 21-megapixel Canon 5d Mark II camera and two PC computers, an entire HMML digitization studio can be assembled for \$7000. If the computers are obtained locally, as is the usual practice, the rest of the system can be shipped to the site in a box measuring 42cm x 42cm x 120 cm. Such a container fits in the back seat of a taxicab and in one instance was hauled on the back of a donkey to a remote site.

The modular nature of HMML's system makes it easy and inexpensive to maintain. All the components are separate, so if something malfunctions, a replacement can be easily sent and swapped with the defective one, minimizing downtime. Any of the components can be upgraded as new equipment appears on the market. When a project is finished, the system packs down into its original shipping box and can be taken to the next project site.

HMML takes advantage of other cost-saving measures. HMML's digitization templates and best practices are stored as pdfs on its website and are freely available to all its studios. Technicians working at HMML projects fill out timesheets, create work reports, and transfer files on free or inexpensive cloud-computing services such as Google Docs and DropBox. Ordinary external hard disk drives are used to send manuscript image data to HMML, where it is copied to a local file server for scholarly use and also backed up to archival data tapes, which are stored at an offsite location.

This practical methodology has worked well for HMML for the past nine years. Over **31,000** manuscripts have been preserved in this manner, with over **130 terabytes** of total data collected. Researchers visiting HMML have unlimited viewing access to over 5 ½ million images of manuscript pages. Those not at HMML can order disk copies of manuscripts for a reasonable fee or take advantage of a new service provided by HMML where up to three manuscripts per scholar are hosted for free online viewing in a password-protected collection in *Vivarium*, HMML's online image viewing website.

3. Closing Thoughts

The Benedictine monks who founded HMML almost 50 years ago had deep instincts of cultural preservation: this is what monks have always done. Those instincts remain among their successors, both lay and monastic, despite all of the changes in technology and scholarship in the past half-century described above. As HMML presents its mission to new audiences, it is important to emphasize that HMML works in service to individual libraries and to the world at large by ensuring that whatever human beings have thought to be worth writing down will not be lost because of war, or lack of resources, or neglect. HMML works with a variety of local partners, ranging from universities with well-developed libraries to very small private collections utterly lacking infrastructure. HMML builds a knowledge base in each place where it works, training local people how to digitize their own manuscript heritage, while serving scholars both locally and internationally. The digitization technicians, who are often young people who would otherwise not have employment, learn the importance of the past even as they use the latest technology to preserve it.

Meanwhile, the larger imperative of cultural preservation is demonstrated in current events. Libraries that HMML has helped to digitize in Syria have suffered looting even as their guardians are forced to flee rising violence. In Jerusalem, HMML works with both Christian and Muslim libraries in the Old City, one of the world's most volatile locations. Young Iraqis displaced from their ancestral homes in Mosul work with manuscripts brought from across the country to the relative safety of the northern region of the country for cleaning, digitizing, and boxing. Archival materials in India are rescued from neglect or destruction, and handled by a team consisting of Hindus, Christians, and Muslims.

HMML is always looking for new partnerships in places where manuscripts are at risk, or where conditions make it difficult for researchers to have access to manuscript materials. We invite questions about our work and about possible collaboration.

Places where HMML has been working since 2003:

- Ethiopia
- India
- Iraq
- Italy
- Jerusalem
- Lebanon
- Malta
- Romania
- Syria
- Turkey
- Ukraine

4. Figures



Figure 1. Digital Single-Lens-Reflex cameras such as the Canon EOS 5D are the central technology in HMML's manuscript imaging methods. They are portable, rugged and relatively inexpensive compared to other technologies.



Figure 2. Microfilming manuscripts in the 1960's. Specialized equipment was bulky and heavy; manuscripts were pressed under glass for photography.



Figure 3. Components of HMML manuscript imaging system ready to pack into oblong shipping box for sending to digitization site. Imaging studio can be shipped in a box measuring 42 x 42 x 120cm.

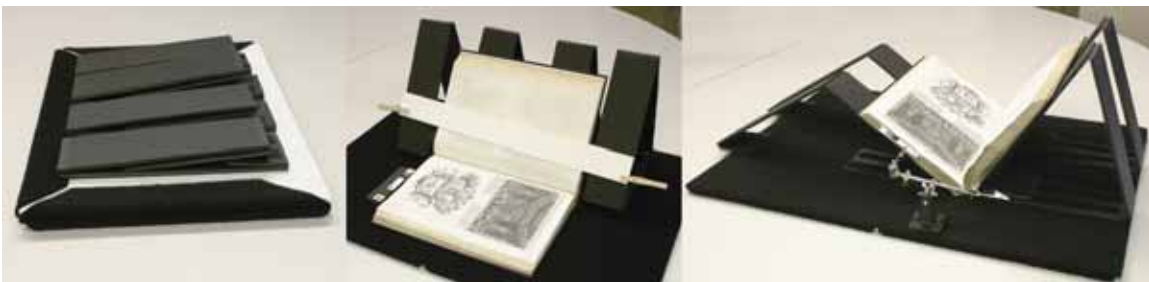


Figure 4. HMML book cradle shown folded for transport, set up for vertical copystand use, and for oblique photography using a tripod.

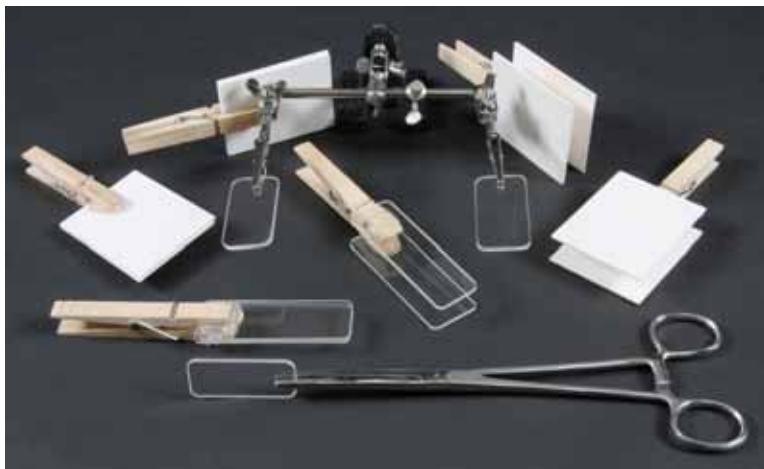


Figure 5. Homemade clips and hold-down tools used by HMML's imaging technicians for working with manuscript codices. Designed in-house in response to feedback from technicians, they are simple, inexpensive and made from readily-available materials.



Figure 6. Manuscript photography at HMML project in Northern Iraq. Manuscripts are being gathered and transported to this studio from all over Iraq. When work is complete, the books are returned to their owners cleaned, boxed, and with copies of the digital data for their own use.



Figure 7. Technician in northern Iraq cleans a manuscript prior to photography.

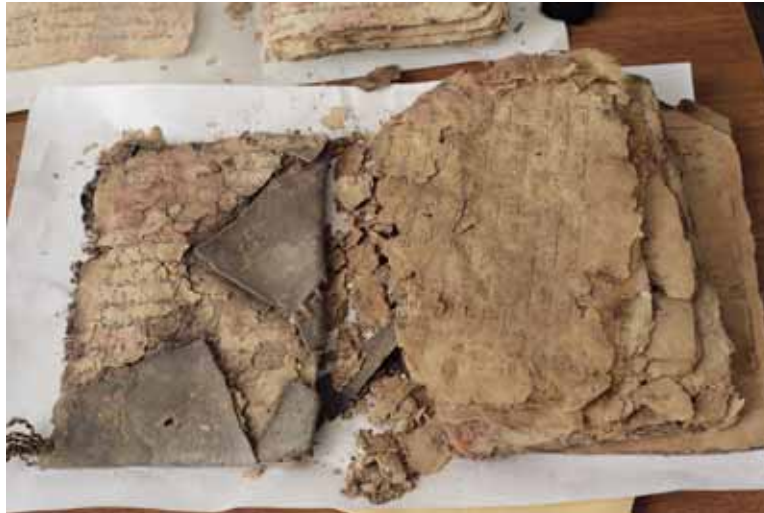


Figure 8. With extremely damaged manuscripts, it must be determined whether successful digitization is possible without destroying the object. HMML's technicians are trained to err on the side of preserving the physical item, as they are often important cultural objects to the communities that own them.



Figure 9. Illumination image from Syrian Orthodox manuscript in Turkey provides a useful comparison between traditional black and white photography and digital colour photography. Colour photography allows researcher to discern image details to a much higher degree.



Figure 10. Prototype solar power system for HMML preservation projects in locations without electricity. Three-panel system can power a laptop computer, external hard drive and camera battery charger. Items fit into carrying case for easy transport.

Plenary 3 Keynotes

Keynote: Challenges for the Preservation of Audiovisual Documents

A General Overview

Dietrich Schüller

Phonogrammarchiv, Austrian Academy of Sciences

Abstract

Unlike text documents, which represent human thoughts in the form of written language, audiovisual documents are representations of physical reality: light and/or sound. For their recording specific carriers and equipment are needed, and reproduction, except for photographs, also depends on the availability of dedicated equipment.¹ Generally, audiovisual data carriers are less stable than traditional text documents, which has already led to deplorable losses, specifically of the early film heritage. But also modern carriers, specifically magnetic tapes, are prone to deterioration. The other threat is the fading of replay equipment, as technical development of audio and video has led to ever shorter life cycles of dedicated audio and video formats, leaving even well-preserved carriers as useless orphans. Recently, classical film preservation has also come under threat as production of traditional film materials is fading, because of the fast spread of digital film projection. Audiovisual documents can only be preserved by content preservation through subsequent digital (=lossless) copying from one IT preservation platform to the next. Analogue contents have to be digitized first. For audio, this shift of paradigm from carrier to content preservation happened already around 1990. The paper explains the technical framework of digitization and digital long-term preservation, analyses the specificities of the various creative sectors of audiovisual documents, surveys the global situation with a special view on developing countries, and summarises the strategic challenges to preserve these documents in the long-term. As a keynote, all attempts will be made to set the framework for all other presentations in the field that will be offered.

Author

Dietrich Schüller, a specialist in audiovisual preservation and former director of the Phonogrammarchiv of the Austrian Academy of Sciences, is Chair of the Austrian National Committee for the Memory of the World Programme. He has worked as a consultant to a number of audiovisual archives world-wide, partly upon request by UNESCO, recently engaged in cooperative projects in Albania, the Philippines and Ethiopia. He is member of the IASA Technical Committee (Chair 1975-2001) and has been actively engaged in the Memory of the World Programme of UNESCO ever since its beginnings, specifically in its Sub-Committee on Technology (Chair 1993-2007). He is also Vice-President of the Intergovernmental Council for the Information for All Programme (IFAP) of UNESCO and Chair of its Working Group on Information Preservation. He is author of numerous publications on audiovisual preservation and a lecturer at several Austrian Universities. He is also engaged in training seminars in Europe and abroad, more recently in Mexico, the Caribbean, China, the Philippines, Singapore, Central Asia, Ethiopia, and Brazil.

It is an almost trivial statement that audiovisual contents form an ever increasing part of political, cultural, educational, and scientific communication. Audiovisual documents have therefore been called the documents of modernity: Without them, the understanding of contemporary history would be utterly imperfect, arts and cultural communication would miss a substantial part of their modern ways of

¹ We have to decide whether or not photographs should be dealt with here or separately.

expression, and intercultural understanding would remain highly rudimentary. This situation is specifically supported by the development of internet over the past years.

Audiovisual recording technologies and their documents, available since the 19th century, have added a substantial extension to human communication and information. While text documents, available for more than 5000 years, transport language and therewith human thoughts, audiovisual documents are equivalents of physical reality: namely sound and/or light.

This fundamental difference is also the reason for their different structural vulnerability: Language, and texts, have a high degree of redundancy. A speck of mould or a small physical damage to a document generally leaves a written text intelligible. With audiovisual documents, however, each detail is potential information, so a similar damage will lead to a more significant loss of content, or may even render the carrier irretrievable.

Unfortunately, this structural vulnerability is accompanied by physical and chemical instability of carriers. While papyrus or parchment has survived for millennia, and acid free paper for centuries, audiovisual data carriers are less stable than traditional text documents. This has already led to deplorable losses, specifically of the early film heritage. But also modern carriers, specifically magnetic tapes, are prone to deterioration.

Limited life expectancy is additionally aggravated by the dependency of audio and video documents on replay equipment. Over the past five decades technical development has lead to ever shorter live time cycles specifically of video formats and their sophisticated replay equipment. Soon after formats became obsolete, the industry withdrew from spare part supply and service support. Ultimately, recording, postproduction and archiving has changed from carrier based dedicated formats to true file formats and the computer world. for audio already in the later 1990s, and for video more recently. This has left great stocks of mainly magnetic tape carriers behind, difficult to replay because of the ever increasing lack of replay equipment. There is agreement that the time window left to keep these machines in operable condition is not more than 15 years, if at all.

This development has been foreseen already around 1990 by audio archivists. In view of the limited life expectancy of carriers, and limited availability of replay equipment, it had become clear that the classical paradigm of museums and archives, namely to preserve the original objects or documents placed in their care, would ultimately be in vain. Long-term preservation has to concentrate on the content by extracting the signals from the original carriers, by digitising them, and by migrating these digitized contents losslessly from one IT preservation platform to the next. At the time, this change of the preservation paradigm was met with scepticism from conservative corners. But already in 1992 the ARD, the Working Group of German Public Broadcasters, started to develop digital mass storage systems which store the master files, control their data integrity, refresh data if needed, and organise migration to the next storage platform with a minimum of manual input. Beyond this automation of the preservation process, the driving force behind the adoption of this new methodology was the remote access to archival holdings, which since has indeed substantially changed the workflow of radio production and transmission.

National audiovisual archives of wealthy countries soon followed this example, and around 2000, after a significant drop of IT component prices, small scale approaches came within financial reach of research archives, including several in developing countries.

Video followed behind audio preservation, mainly because of the size of video files, which are up to 50 and more times bigger than audio files. Today, however, big institutions can afford so-called *petastores* to handle their great amounts of digital data.

Most recently, film preservation, which is in need of even greater storage capacities, is also entering digital preservation methodology. The driving force here is the ever expanding digital film production and projection, which leads to a rapid fading of classical photo-chemical (analogue) film, which is needed for traditional film preservation.

While with text documents predominantly a tool for democratising access, digitization is an inescapable measure for the preservation of audiovisual documents. Because of its implementation already in the early 1990s, and the great sizes of audiovisual files—as compared to text files—audiovisual archives have played a pioneering role in digital long-term preservation.

The basic principle of *archiving* is to preserve the information we have stored for future access, *complete* and *unmodified*. Should the concept of content migration work in accordance with this basic principle, several conditions must be in place, and several challenges have to be met:

Contents have to be completely extracted from original carriers, which is a challenge specifically with historical carriers and formats. Generally, the use of modern equipment, modified to the historical parameters, is to be preferred over the use of historical equipment, which would add, due to its technical limitations, unwanted replay distortions to the replay signal. With modern equipment, however, historical recordings normally can be retrieved in a better way than at the time of their production. Generally no significant further improvement of signal extraction may be expected, but there are some niches for potential further development, e.g., optical replay of mechanical carriers.

The greatest problem in the field of signal extraction is, first of all, the mere availability of adequate modern replay equipment, specifically for the great variety of video formats. Further the challenge of maintaining that highly dedicated equipment in operating condition in a world of fading spare part and service supply. To date, audiovisual archive technicians have increasingly to keep up the skills of maintaining equipment, previously carried out by the manufacturers, and this trend will progress irreversibly.

Digital audio and video recording started already in the 1980s. The safeguarding of digital recordings, however, is not much easier than digitising analogue recordings. Practically all these early so-called single carrier formats are obsolete, except CDs and DVDs, and equipment is equally as scarce as that for analogue originals. Archival standards call for the conversion of those early recordings into true file formats, by keeping the original resolution. With data reduced originals the original encoding shall be maintained, if possible.

An often highly underrated aspect in the planning of digitization projects is the time factor. The time needed to extract contents from digital contents exceeds by far the playing time of contents. Cleaning and loading of carriers, checking and compensating for recording errors, production of metadata, to mention but few of the operations needed, extend the operation to about three times of the playing time for audio, and more for video. Damaged or otherwise problematic carriers need considerably more. In order to digitize the holdings of broadcast audio and video collections, software supported methods have been developed that allow a computer controlled transfer of up to five carriers of the same type by one operator simultaneously. This *factory transfer*, however, can only be applied to materials that have been recorded under standardised and controlled condition, which is generally the case with the greater part of stocks of radio and television archives. Collection that hold originals of different origin, often produced under technically uncontrolled conditions, such as national archives, generally have to employ manual transfer with all time consuming consequences.

On the side of digital preservation the predominant problem is the choice of digital resolution to adequately represent the analogue original. Since digitization has started in the 1990s, there is a trend to

ever higher digital resolutions for audio, video, and now for film. This is supported by the advancement of conversion technology (A/D converters, scanners), along with an ever decreasing price of storage space, which, for example, has dropped to one tenth over the past five years, with a tendency to drop further, probably at a slower rate. A significant advantage of high resolution concerns restoration, which is the more efficient, the more information is available on the unwanted artefacts.

In the earlier days of audio and video digitization, however, storage costs have been a considerable factor. Therefore, data reduced, also called data “compressed” formats have been employed for archival files, which predominantly happened in the work of television archives. Also, in order to make audiovisual content accessible over the internet, strong data reduction algorithms have been developed, which have made low quality consumption of audiovisual content widespread. Data reduction (“compression”) is eliminating data and hence cannot be reversed. An MP3 file does not fully represent an analogue LP signal, neither does a DVD represent the information contained in a film. The use of lossy “compression” in archiving is, therefore, incompatible to basic archival principles. Data reduction also creates disadvantages for the further use of such signals.

This has more recently led to the acceptance of archival standards proper for institutions which are predominantly following commercial interests, beyond archival obligations in the narrower sense. Lossless compression, however, e.g., MJPEG 2000, is compatible with archival standards, and is presently widely employed, assisting to reduce significantly the costs of uncompromising high resolution video and film archiving. It must be noted that, while problems of standard definition (SD) video archiving are consolidating, upcoming HD and 3D video constitutes a new challenge in terms of increasing storage space requirements.

The choice of openly defined, non-proprietary formats is another important factor of digital archiving. In contrast to digital text and image preservation, which has struggled with early, and ever changing consumer formats, audiovisual archiving is practically unaffected by this problem. Mainly due the early initiative of the European Broadcasting Union (EBU) the WAVE format has become the de-facto standard for audio. Format consolidation of video files is presently happening, while digital film archiving is following the standards set by D-Cinema. Dissimilar to other branches of digital preservation, format obsolescence will not realistically become a problem in the mid-term.

Another aspect peculiar to digital long-term preservation of audiovisual documents is their considerable amount of data. While text and image collections have started by widely using recordable optical discs as storage media, professional audio archiving started out by using professional IT components. Only small, typically research collections, have used recordable optical discs and other digital consumer formats for a short period, but from around 2000 small scalable solutions employing IT equipment have been developed. Small collections manage to some extent the manual control of their holdings, but as the amount of data general rises quickly beyond 10 Terabyte, software controlled storage management becomes indispensable. Originally, such software was a considerable cost factor, but meanwhile, open source software, also developed with support from UNESCO, is available, thus assisting specifically small collections in developing countries.

The standards and guidelines for audiovisual archiving are set by professional non-governmental organisation such as IASA (International Association of Sound and Audiovisual Archives), FIAF (Fédération Internationale des Archives du Film), SMPTE (Society of Motion Picture and Television Engineers) and others. Guidelines are also produced and published by international projects such as the

European Commission funded PRESTO family projects,² or TAPE,³ or, as a national example, the (US) National Digital Information Infrastructure and Preservation Program (NDIIPP). led by the Library of Congress

As consequence of the above explained problems and required procedures, long-term preservation of audiovisual documents requires specialised equipment and high level expertise for its maintenance in a world of fading industrial support, to retrieve the contents from their originals and—like all other sectors of digital long-term preservation—ongoing personal, strategic and financial engagement to keep digital files alive. It is obvious that effective, professional work can only be achieved when expertise and funds are concentrated in institutions that hold critical masses. Therefore, at the start of audiovisual preservation planning, a critical review of the collection policy is needed. Often, audiovisual collections have been acquired more or less coincidentally as part of conventional libraries and archives. Transfer of such holdings to more specialised institutions should be considered. Mere possession should not be a guiding principle, but the optimisation of access to the documents. Several cooperative models have been developed, some of them on a national basis, like in Switzerland, others within institutions, as recently at the Indiana University Bloomington, or at the University of the Philippines, where central transfer laboratories are established for the holdings of the institutes, while the preservation of digital files is entrusted to the computer centre of the campus. Finally, the employment of private vendors for digitization, sometimes also for digital long-term preservation, is spreading. While fairly well introduced in North America, outsourcing of archival work is presently gaining popularity also in other parts of the world.

More than 20 years after the change of paradigm from carrier to content preservation, the technical and methodological way to proceed into the future is clear: audiovisual preservation by digitization and permanent future migration of digital files is possible and has become a widely applied routine. It is, however, strategically and financially demanding, and its effective application for safeguarding of the *global* holdings must be seen with concern.

The world-wide audio and video collections are estimated to amount to 200 million hours. This figure may be fairly realistic, but it has never been critically examined for the multiple copies it contains.

A great portion is held by the broadcast radio and television archives. Generally the archives of (former) public broadcasters are much bigger than those of commercial stations and many of their holdings are unique documents of considerable historical and cultural value. The other greater part of audiovisual documents is held by national audiovisual archives, or by audiovisual departments of national, regional or municipal libraries and archives. These typically contain the products of the (national) audiovisual industry, but also unique material related to the history and culture of the respective countries or regions.

It can be assumed that such holdings of wealthy countries will be preserved, at least in a selective way. Less optimistically must be seen the holdings of less wealthy and developing countries. Here,

² Presto started 2000 with emphasis on preservation problems of radio and television archives <http://presto.joanneum.ac.at/projects.asp>; the later Presto projects aimed at making broadcast archive technology also available for other audiovisual archives:

PrestoSpace <http://digitalpreservation.ssl.co.uk/>;

PrestoPrime <http://www.prestoprime.org/project/public.en.html>; and

PrestoCentre <http://www.prestocentre.org/>.

³ TAPE (Training for Audiovisual Preservation in Europe) 2004-2008, aimed specifically at small audiovisual collections in non-specialised institutions, such as libraries, conventional archives and museums: <http://www.tape-online.net/>.

CCAAA, the umbrella organisation of audiovisual archives association is trying to assist through its Archives@Risk Programme. However, in the global business driven atmosphere, accompanied by a financial crisis, it is increasingly difficult to raise funds for culture and humanities.

From the point of UNESCO, the greatest present threat concerns the many small and hidden audio video documents, which have been recorded worldwide by academic and cultural institutions of languages, traditional music, dances, theatres, rituals and cultural events. They are the documents proper of cultural and linguistic diversity, one of the main objectives of UNESCO. Probably 80% of these important cultural holdings are outside archival custody, held by notoriously underfunded institutions, sometime even by private researchers, partly even unaware of the imminent threat, and, even if so, generally unable to raise the funds for their safeguarding. Many of these documents contain languages and cultural expressions which have meanwhile vanished, or at least significantly changed. Though it was not possible to keep these traditions alive, all should be done to at least preserve their documents. This situation is specifically aggravated by the limited time window of—hopefully—15 years, the time span left of the availability of audio and video replay equipment.

The present situation is without historical precedent: Unless decisive and concerted action is taken, a considerable part of the primary sources that form our present knowledge of the linguistic and cultural diversity of human kind will vanish.

In order to react adequately to this situation, the IFAP Working Group on Information Preservation with its core, the Memory of the World Technical Subcommittee, has proposed to UNESCO to undertake a survey on information preservation, to realistically assess the quantitative dimension of threats endangering the documentary heritage and specifically embark on a mid-term project for the safeguarding of the scattered and hidden documents of cultural and linguistic diversity. These projects had to be postponed, due to the present financial constraints.

However, in view of the limited time window, these projects cannot wait much longer. Specifically delegates from parts of the world rich of orally transmitted cultures may feel challenged to include these urgent projects into the recommendations of this conference to UNESCO.

Select Bibliography

- Academy of Motion Picture Arts and Sciences (Ed.). *The Digital Dilemma 2*. Hollywood, 2012.
<http://origin-www.oscars.org/sciencetechnology/council/projects/digitaldilemma2/download.php>.
- Bradley, Kevin. “Risks Associated with the Use of Recordable CDs and DVDs as Reliable Storage Media in Archival Collections - Strategies and Alternatives.” Paris: UNESCO, 2006.
 web version: <http://www.unesco.org/webworld/risk>.
- Bradley, Kevin. “Towards an Open Source Repository and preservation System. Recommendations on the Implementation of an Open Source Digital Archival and Preservation System and on Related Software Development.” Paris: UNESCO, 2007.
<http://www.unesco.org/webworld/en/mow-open-source>.
- Edmondson, Ray. *Audiovisual Archiving: Philosophy and Principles*. Paris: UNESCO, 2004.
 (CI/2004/WS/2). <http://unesdoc.unesco.org/images/0013/001364/136477e.pdf>.
- IASA Task Force on Selection, Majella Breen et al. “Selection criteria of analogue and digital audio contents for transfer to data formats for preservation purposes.” International Association of Sound and Audiovisual Archives (IASA), 2004. web version: <http://iasa-web.org/de/task-force>.

- IASA Technical Committee. *The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy*, edited by Dietrich Schüller (= IASA Technical Committee - Standards, Recommended Practices and Strategies, IASA TC-03), Version 3, 2005. web version: http://www.iasa-web.org/downloads/publications/TC03_English.pdf. Also available in French, German, Italian, Russian, Spanish, Swedish, and Chinese
- IASA Technical Committee. *Guidelines on the Production and Preservation of Digital Audio Objects*, edited by Kevin Bradley (= IASA Technical Committee - Standards, Recommended Practices and Strategies, IASA TC-04), 2nd ed, 2009. web version: <http://iasa-web.org/audio-preservation-tc04>.
- Henriksson, Juha, und Nadja Wallaszkovits. "Audio Tape Digitisation Workflow. Digitisation workflow for analogue open reel tapes." March 2008. <http://www.jazzpoparkisto.net/audio/>
- Klijn, Edwin, and Yola de Lusenet. "Tracking the reel world. A survey of audiovisual collections in Europe." January 2008. http://www.tape-online.net/docs/tracking_the_reel_world.pdf
- Schüller, Dietrich. "Audiovisual research collections and their preservation." April 2008. http://www.tape-online.net/docs/audiovisual_research_collections.pdf.
- Schüller, Dietrich. "Socio-technical and Socio-cultural Challenges of Audio and Video Preservation." *International Preservation News* 46 (December 2008): 5-7. <http://archive.ifla.org/VI/4/news/ipnn46.pdf>.
- Schüller, Dietrich. "Video Archiving and the Dilemma of Data Compression." *International Preservation News* 47 (May 2009): 5-7. http://www.ifla.org/files/pac/IPN_47_web.pdf.
- Schüller, Dietrich. "Audio and Video Materials in Tropical Countries." *International Preservation News* 54 (August 2011): 31-3. http://www.ifla.org/files/pac/ipn/IPN_54def.pdf.
- Wright, Richard. "Broadcast Archives: Preserving the Future." http://presto.joanneum.ac.at/Public/ICHIM%20PRESTO%2028_05_01.pdf.

International Perspectives and Cooperation

One Year of Efforts for Digital Preservation at FAO

Claudia Nicolai,¹ Rachele Oriente¹ and Fernando Serván²

¹Knowledge and Information Management Officer, Food & Agriculture Organization of the United Nations

²Chief, Internet and Internal Communications, Food & Agriculture Organization of the United Nations

Abstract

This paper describes the spectrum of challenges faced and actions taken in the first year of working on digital preservation in the Food and Agriculture Organization of the United Nations (FAO). FAO faced ongoing challenges to find funding; political challenges of getting support from management of the need for digital preservation; cultural challenges of working across silos without clear internal responsibilities to require compliance, a scale of the work that demands curtailing preservation activities to the confines of the possible rather than the ideal, and the legal and normative challenges related to preserving our Web site. The work that FAO began to overcome these challenges includes preparation of documentary groundwork, capacity development in the subject, procurement for web archiving services, and exploration of the benefits of UN inter-Agency work through a survey of digital preservation activities in the UN Agencies.

Authors

Claudia Nicolai is a Knowledge & Information Management Officer in Internet and Internal Communications at the Food & Agriculture Organization of the United Nations in Rome. She has MSc (University of Rome, 1995) in Biology and has worked in research institutions before joining FAO in 1999. In FAO she has been mostly involved in development and maintenance of document workflow, repositories and dissemination through the FAO Web site. She has recently moved to the Internet and Internal Communications unit, where she is responsible of the Digital Preservation activities related to the FAO Website. She has been involved with Digital Preservation for the past year and working with colleague Rachele Oriente.

Rachele Oriente is a Knowledge & Information Management Officer in the David Lubin Memorial Library at the Food & Agriculture Organization of the United Nations in Rome. She has MLS (UBC, 1988) and a MA (UBC, 2003) and has worked in corporate and academic libraries in government, inter-governmental and international organizations and corporations. She has been involved with Digital Preservation for the past year and working with colleague Claudia Nicolai.

Fernando Serván. holds an MSc in Mechanical Engineering (Universidad Católica del Perú) and an MBA (Universidad de Barcelona). He has been involved with the FAO website and related electronic resources since the FAO website started in 1995 and responsible of official multilingual document production until February 2012. Since March 2012 he is the Chief of the Internet and Internal Communications in FAO.

Background

FAO is a knowledge organization with a constitutional mandate to create, disseminate and give access to agricultural information. FAO has an intricate organizational structure with decision-making authority descending through seven departments and eight offices, each with multiple subsidiary offices or divisions and worldwide regional, sub-regional and country offices.

FAO is challenged to ensure that all information providers throughout the organization understand the vulnerability of digital information and commit to safeguarding FAO's digital heritage by following a

comprehensive and standardized approach to management and curation of digital objects. A problem statement was prepared, together with a video and proposals to elicit the management support without which the work cannot succeed, and it was submitted to the relevant governance structures. The UN-wide survey confirmed on the UN system level what the FAO experience has been: lack of funding and of management support are the two most significant obstacles to successful digital preservation work in this type of organizations. Understaffing, multitasking staff and the need for training are challenges connected to the lack of funding.

A workshop with external experts was held in May 2011 and further workshops are planned for 2012. FAO produces a vast amount of information in born-digital and digitized print formats and other media. The scale of the work demands curtailing preservation activities to the confines of the possible rather than the ideal. Delay will mean that the sheer volume of objects that merit preservation will overwhelm resources, resulting in permanent loss. Some resources are more vulnerable to attrition than others: the FAO Web site, for example, changes quickly in content, appearance and structure. The Web site holds vast amounts of valuable and substantive content in millions of pages. Web sites are managed by divisions and departments so control over loss of information as Web sites are updated is a challenge that incorporates the spectrum of economic, political and organizational culture challenges within one activity.

It is foreseen that Digital Preservation practices will commence in the institutional repository (Open Archive) and with Web Archiving in 2013.

1. Introduction

This paper relates the spectrum of challenges faced and actions taken in the first year of working on Digital Preservation in FAO. The last few years have been times of changes and renovation in FAO. The entire Web site is about to be re-organized in order to make information more usable and accessible to a vast audience. The institutional repository for publication has been re-engineered and the Organization is moving towards a clear policy of Open Access in the framework of Open Archival Information System. The need to start thinking about the time dimension of accessibility to FAO materials has made itself known recently. The amount of born digital or digitized information FAO produces is impressive and FAO must overcome the misperceptions that many information producers have: digital objects will simply be there forever or—when they are lost—there was nothing that could be done to prevent loss. The task of implementing Digital Preservation in FAO seems gigantic and the challenges are numerous. The first year of activities shows that, step after step, some results can be achieved.

2. What is FAO?

FAO is a specialized agency of the United Nations. FAO's structure comprehends seven Departments and a decentralized network with regional, sub regional and country offices, which mark FAO's presence in over 130 countries.

FAO was founded in 1943 at Hot Springs, Virginia, United States of America, during the UN Conference on Food and Agriculture. The Conference was organized during the World War II when 44 nations met under strict security measures and committed themselves to founding a permanent organization supporting development in the field of agriculture and devoted to ensure food security. FAO

was formally instituted during the First Session of the FAO Conference, held in Quebec, Canada, in 1945.¹

FAO's mandate is to raise the level of nutrition, improve agricultural productivity, better the lives of rural populations and contribute to the growth of the world economy. FAO is a knowledge organization and regards itself as a source of knowledge and information in the field of agriculture, commodities, nutrition and sustainable development among others.²

To this end, FAO is mandated by its Constitution "to collect, analyse, interpret and disseminate information relating to nutrition, food and agriculture."³ The FAO Knowledge Strategy articulates a need for a preservation strategy and practice. Principle 5 of the Knowledge Strategy, Global Perspective, determines that: "FAO will play a key facilitation role in ensuring the world's knowledge resources are available to those who need it, when they need it and in a format they can access and use."⁴

As a UN Agency, FAO should comply with system-wide statements on purpose or goals. During the 32nd Session of the UNESCO General Conference a Charter on the Preservation of Digital Heritage⁵ was adopted that recognizes that "*digital heritage is at risk of being lost and that its preservation for the benefit of present and future generations is an urgent issue of worldwide concern*," thus reinforcing the need for FAO for preserving and guaranteeing access over time to the organizational digital assets is crucial to FAO.

3. Why is Digital Preservation needed in FAO and why is it a challenge?

The amount of information that FAO produces is impressive and varied. Over a million times a month, someone visits the FAO Web site to consult a technical document, access statistical data or find more information on FAO's projects, maps and multimedia.

FAO is also responsible of publishing hundreds of newsletters, reports and books, distributes several magazines, creates numerous CD-ROMS and hosts dozens of electronic discussions. It collects Country's data on Agriculture since 1945 and in some cases even older data. Access to this information is of uttermost value for agricultural institutions of FAO's member countries. FAO's collections of statistical data are accessible via Internet in FAOSTAT,⁶ the statistics database.

The quantity and quality of this information outputs requires care and custodianship at every stage of its lifecycle. Although corporate repositories exist, a considerable amount of resources slips through the cracks of the systems and is neither archived nor recorded, let alone adequately preserved.

These resources are the more prone to Digital Preservation related risks. Project documentation can be a good example, since many project related documents are very likely to disappear after the project closes. In some cases the documents are uploaded into a project Web site and if the Web site is not preserve neither the documents will be.

¹ FAO: its origins, formation and evolution 1945–1981, accessed August 31, 2012, <http://www.fao.org/docrep/009/p4228e/P4228E00.htm#TOC>.

² FAO online, knowledge management and more, accessed August 31, 2012, <ftp://ftp.fao.org/docrep/fao/011/i0765e/i0765e07.pdf>.

³ FAO Constitution, article 1, accessed August 31, 2012, <http://www.fao.org/docrep/meeting/022/K8024E.pdf>.

⁴ FAO. *FAO Knowledge Strategy* (March 2011): 6, http://www.fao.org/fileadmin/user_upload/capacity_building/KM_Strategy.pdf.

⁵ UNESCO 32nd Conference, accessed August 31, 2012, <http://unesdoc.unesco.org/images/0013/001331/133171e.pdf>.

⁶ FAOSTAT, accessed August 31, 2012, <http://faostat.fao.org/>.

FAO has a designated audience that is extremely wide and diversified, including scholars, global agricultural community, NGOs, journalists, academic and research institutions, policy makers and the general public from all over the world.

The benefits of Digital Preservation would extend beyond FAO's own outputs. Member countries will benefit not only from accessing FAO's work, but because FAO will include some selected materials produced by our member countries that they may be unable to preserve, such as paper copies of statistics for FAOSTAT.

The motivation to start Digital Preservation activities in FAO is thus very strong notwithstanding the complexity of the task.

4. One year of Digital Preservation in FAO

May 2011 can be considered as the starting point for Digital Preservation activities in FAO.

A Digital Preservation and JHOVE2 Next-Generation Characterization Workshop took place in the Organization's premises. The objective was to provide internal participants with a general understanding of Digital Preservation concepts, goals and practices. The last two days of the event were focused instead on the knowledge of JHOVE2's capabilities and architecture and its potential role in local workflows.

Following that first workshop and during the entire first year of work FAO faced ongoing challenges to allocate funding; internal challenges of getting support from management of the need for digital preservation; cultural challenges of working across silos without clear internal responsibilities to require compliance, a scale of the work that demands curtailing preservation activities to the confines of the possible rather than the ideal, and the legal and normative challenges related to preserving our Web site.

The work that FAO has begun to overcome these challenges includes preparation of documentary groundwork, awareness raising, preliminary assessment of FAO's digital assets, capacity development, initial procurement for Web archiving services, and exploration of the benefits of UN inter-Agency work through a survey of Digital Preservation activities in the UN Agencies.

A problem statement on Digital Preservation was prepared in order to inform FAO management in critical units of the existing threat of information loss and also in order to have the preservation need formally acknowledged.

The awareness raising exercise has involved many other entities, including the technical departments who are the main producers of information, the IT and governance units who are mostly concerned with procedures and rules across the organizations and the newly established Internet and Internal Communications team that plays a crucial role in the implementation of FAO's Internet and communication strategy.

Gaining official acknowledgement of the Digital Preservation needs in the organization has been a huge challenge, especially in times of severe budget cuts.

There were three main arguments that supported the Digital Preservation cause: credibility, cost efficiency and identity.

1. *Credibility*: Guaranteeing long-term access and authenticity of FAO digital heritage is crucial to FAO's credibility as a knowledge organization.
2. *Cost efficiency*: The cost of Digital Preservation at the production stage is much less than retrospectively attempting to rescue digital objects that are no longer accessible or resuscitating objects with faulty metadata. So far FAO has invested in the initial field work, research,

intellectual analysis and infrastructure to create information products; preservation will ensure that this investment will endure and yield long-term returns by long-term utility. To successfully manage an information product throughout its full life cycle, preservation actions must be incorporated into each stage of production, use and storage.

3. *Identity*: FAO's identity is represented in its organizational memory. By preserving its tacit knowledge that is encoded in the thousands of born-digital and digitized publications, project documentation, series of statistics spanning over almost seven decades, information systems and Web pages, FAO is actually preserving its own identity as an organization. If FAO was not to preserve them the organization would be comparable to a human being progressively losing his/her long-term memory. It is also worth noting that FAO knowledge is a key element of the collective memory of the world as FAO holds information with a vast time, language and geographic coverage - in some cases this information is not available anywhere else, not even in the countries that actually produced it.

Although in the last few months FAO has only taken the initial steps towards the implementation of Digital Preservation at organizational level, these steps seem to be the right ones.

The Digital Preservation problem statement was formally initiated by FAO the Office of Knowledge in October 2011 and was also distributed to and acknowledged by Communication and IT functions of the Organization. Some resources were allocated to support the UN survey on Digital Preservation and to organize two workshops: one focusing on the FAO Open Archive for publications and the other on following up to the UN survey and engaging a wider UN audience on the subject of Digital Preservation.

The UN survey was open for data collection from December 8, 2011 to January 15, 2012 and was sent to a list of 44 individual contacts in the library, archives, records management and information technology areas of 25 UN Agencies.

The survey responses indicate both recognition of the timeliness for implementing Digital Preservation and that expertise must be developed within the organizations. The survey confirmed on a UN system level what the FAO experience has been: lack of funding and of management support are the two most significant obstacles to successful Digital Preservation work. Understaffing, multitasking staff and the need for training are challenges connected to the lack of funding. The information gathered from this survey has informed FAO's work, helped plan a workshop for UN participants that is scheduled for October 22-23, 2012 and to tailor the sessions to the needs of UN Agencies. The event will be linked to the organization of the UN-LINKS meeting for late October 2012.

March was a month particularly favorable to Digital Preservation activities in FAO. A multi-day workshop with external experts from the Technical University of Vienna was held in March 2012. The workshop was aimed at engaging relevant FAO actors with the support of international Digital Preservation experts. The workshop's topics included the foundations of Digital Preservation, case studies and was meant to encourage stakeholders' discussions. The workshop was specifically focusing on planning using Plato-Preservation Planning Tool.⁷ The planning exercise was focused on the Digital Preservation needs of the FAO Open Archive, which is the official archive for FAO documentation and also a tool in the implementation of digital preservation activities. The planning exercise was carried out using the different PDF versions collections and included evaluation of the possible alternative formats for migration, based on the selected significant properties of the digital objects. As preparatory work for the

⁷ Plato: the Preservation Planning Tool, accessed August 31, 2012, <http://www.ifs.tuwien.ac.at/dp/plato/intro.html>.

workshop PDF and HTML collections, including high resolution PDF were assessed and characterized using Format Identification Tools System. The outcome has been used as a starting point and baseline study for the definition of a long-term preservation plan for the Open Archive digital collections.

The FAO Open Archive is in its final implementation phase and release is foreseen during 2013. The possibility to include the FAO multimedia collections in the Open Archive is being considered, once the Open Archive system is finalized.

The proposal to define and implement a Digital Preservation strategy at Organizational level—as a way to safeguard authentic and access to the organizational digital heritage—was identified as one of the priority for extra budgetary funding allocation for the upcoming biennium (2013-2014).

The proposal to define and implement a Digital Preservation strategy for the FAO Web was incorporated as part of the overall proposal for a new FAO Web strategy. This proposal was jointly supported and submitted by different FAO units. Procurement for the identification of suitable service provider for Web archiving services has started and the service is foreseen to be active by the beginning of 2013. It is worth noting that Web archiving also officially appears as an activity line in the FAO Programme for Work and Budget for 2012-13.

4.1 A Major Challenge: Preserving the FAO Web

The FAO Web site is a pertinent example of subject, legal and normative imperatives for creating a Digital Preservation strategy. Moreover, a recently published UN report⁸ refers to access to information available on the Internet as a human right. Preserving and granting long-term access to FAO Web content is of particular relevance.

It's not surprising that the FAO Web has been identified as a priority for preservation. At the same time it is also worth considering that the preservation of FAO Web is also a major challenge because of its complexity and its vast, variegated and multilingual content.

The first appearance *fao.org* dates back to 1994. At that time the FAO Web site was considered an experimental innovation activity conducted by the IT and General Information units. In FAO, as with the rest of the world, the attention toward this new communication channel grew very rapidly. Since the beginning, FAO management gave great emphasis on providing information on FAO via the Internet. By mid-1996, with the launch of Cerestronic (a news-driven FAO homepage) FAO already had a Web presence covering most of the technical areas in its mandate.⁹

The FAO Web site changes quickly in content, appearance and structure and holds an incredible amount of content. Since its first appearance the FAO Web site has grown exponentially in terms of size, to the point where there are now more than 4 million files on *fao.org*. A preliminary analysis of *fao.org* already identified over 390 applications and databases distributed in more than 800 different Web sites containing a plethora of pages, documents, multimedia, electronic for a, statistical data and maps.

In the last 15 years the FAO Web site has also exponentially grown in usage: it is one of the most visited in the UN system and receives one million visits a month, with peak periods such as October 2011, when over 5.4 million visits were registered.

⁸ Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue, accessed August 31, 2012, <http://www.un.org/Docs/journal/asp/ws.asp?m=A/HRC/17/27>.

⁹ FAO Conference. PROGRAMME of WORK and BUDGET 2000-1, accessed August 31, 2012, <http://www.fao.org/docrep/X2500E/X2500e15.htm>.

The FAO Web site changes rapidly and is a reflection of the changes in the organization and in the world, every time the Web site changes thousands of Web pages and documents change too. Content is created and then is lost forever as pages, documents, files or other media are removed from the Web in order to update it. Volatile content out of www.fao.org is worth preserving.

In an attempt to create a visual presentation on the FAO Web and to advocate for the need to preserve it, we collected a selection of the FAO home pages preserved in the Wayback machine of the Internet Archive. The video produced was quite a nice way of visualizing in a couple of minutes how FAO Web site has evolved in the last 15 years.

The UK Web Archiving Consortium (UK WAC) estimates the average life of a UK Web site is the same as a housefly—about 44 days. For FAO in 1996 the lifespan of the home page was 47 days.

In November 1996 FAO Web site was still only available in English, French and Spanish (Figure 1). In the lifespan of a housefly (plus three days!) the Arabic speaking world was able to access FAO's information in its own language (Figure 2). The preparatory work took far more than a few weeks. Of course, information had to be translated but also technical problems related to the encoding of Arabic were to be solved, Unicode days were still to come.



Figure 1. FAO homepage (November 4, 1996).



Figure 2. FAO homepage (December 21, 1996).

Multilingualism is a considerable added value of the FAO Web site. Access to information and knowledge in all official languages has always been strongly supported by FAO management. Appearance of the Chinese version of the Web site dates back to 2000, while Russian version appears in 2009 soon after the inclusion of Russian as FAO's official language.

Specific sections of the FAO Web have legal obligations to preserve web content over a period of time—for instance, the International Plant Protection Convention, a web site hosted by FAO, has a legal obligation to preserve any published content for 10 years (www.ippc.org).

All elements were there to support the evidence that Digital Preservation of the FAO Web was needed.

The Web Strategy implementation will ultimately result in a new and improved FAO Web site along with guidance and recommendations for web content publishers. The definition of a Digital Preservation strategy for the FAO Web site, the detailed planning of Digital Preservation activities, and the setting up of Web Archiving services, have been identified as a key elements of success for the Web Strategy implementation.

Beside the evident value of preservation an additional reason to include Digital Preservation in the new Web strategy is that not all FAO Web sites are compliant with FAO Web publishing standards. A Digital Preservation strategy can help entailing standards for Web sites so that they are compliant with FAO web publishing standards in order to preserve them.

The definition of a Digital Preservation strategy and plan are among the first activities foreseen for the implementation of the web strategy.

Another activity that is relevant to the preservation of the FAO Web is the creation and maintenance of an “Inventory of FAO Web sites,” which will also include project sites and non-FAO domains hosted by FAO. The entire fao.org will be fully assessed, each Web site will have an ID card, containing—among other—metadata related to its preservation. All Web sites are going to be scored against a set of quality criteria, and according to the results of the scoring exercise a plan to sunset Web sites based on scoring criteria will be produced. Besides, a list of Web sites that need to be archived on a regular basis will also be prepared.

It is also foreseen that all FAO users will be able to request archival services for the Web site they are responsible for. The FAO Web team will be responsible of granting or denying these services and will determine the archival schedule.

Moreover some criteria have already been outlined in order to help identify valuable Web content to be archived:

- FAO Web sites where we have legal obligations to preserve any published content over a period of time;
- Obsolete FAO Web sites or those that are not regularly updated and should be removed from the web, will still be accessible but in the right context, as archived information;
- FAO Web sites with volatile content;
- Selected external sources of information that are linked for FAO web site. For example the FAO Country Profiles, in order to take a snapshot of the information on a given country at a given date;
- Information related to FAO’s work that is available on the web, possibly including FAO related social media accounts.

The requirements to identify a service provider for Web archiving have already been defined. A tender is about to be carried out in order to identify the most suitable service provider.

FAO is clearly only about to step into a new incredible challenge, with considerable opportunities to improve the way its Web site is governed managed and preserved.

5. Conclusions

FAO is expert in the creation and initial dissemination of knowledge but has only just started to address the preservation of that knowledge with sufficient priority. The continued absence of policy, process and procedure to ensure long-term access to information would be a tacit denial of responsibility for stewardship of the information FAO has created.

Loss of organizational heritage would hinder agricultural development and food security progress and hamper the creation of new knowledge in the global agricultural community.

Delay would mean that the sheer volume of objects that merit preservation will overwhelm resources, resulting in permanent loss. For the works that are lost, scientific and technical research would have to be replicated at a higher cost than that of preserving the information now.

The need to define and implement a Digital Preservation strategy in FAO is clear, as it will put into operation the preservation procedures and activities that will pre-empt these threats.

With a proper Digital Preservation strategy in place, not only FAO stakeholders but also member countries will benefit from long-term access to FAO documentation regarding their countries. A Digital Preservation strategy will enable and support FAO in its work to end hunger through information-based solutions, by ensuring perpetual access to FAO's intellectual outputs.

After slightly more than a year of activities FAO has achieved some significant objectives. The dimensions of complexity at FAO have been political, economic and cultural.

FAO has obtained management support in the units covering the areas of Knowledge, Communication and IT. The support was enough to have funds allocated, to start activities for the preservation of the FAO Web and to see the need for an organizational Digital Preservation strategy acknowledged as a priority for the upcoming biennium.

The challenge still ahead, with the greatest impact on Digital Preservation work at FAO, is to consolidate these initial political and economic achievements by gaining full support from all management over the long-term. The cultural challenge is still outstanding as FAO has to make sure that all information producers throughout the organization understand the vulnerability of digital information and commit to safeguarding FAO's digital heritage by following a comprehensive and standardized approach to management and curation of digital objects.

Activities related to the preservation of the FAO Web have already started- and that in itself is currently an enormous new challenge and opportunity—and Digital Preservation practices are planned to commence also in the institutional repository (Open Archive) in 2013.

Archiving the World's E-Journals

The Keepers Registry as Global Monitor

Peter Burnhill,¹ Françoise Pelle,² Pierre Godefroy,² Fred Guy,¹ Morag Macgregor¹ and Adam Rusbridge¹

¹Based at EDINA, the UK/JISC national data centre at the University of Edinburgh, Scotland UK, <http://edina.ac.uk>; ²Based at the ISSN International Centre in Paris, France.¹

Authors

Peter Burnhill is Director of EDINA, the JISC National Data Centre at the University of Edinburgh in the UK. A past president of IASSIST, Observer with the ISSN Network and the first director of the Digital Curation Centre, Peter is active nationally and internationally in strategy, design and build of digital infrastructure.

Françoise Pellé is Director of the ISSN International Centre, the intergovernmental organization in charge of the management of the ISSN system. Françoise is also the chair of ISO / TC 46, the technical committee of ISO devoted to standardization in information and documentation, and member of IFLA, the International Federation of Library Associations, where she is active in the serials section and committee on standards.

Pierre Godefroy is Project Manager at the ISSN International Centre, the intergovernmental organization in charge of the management of the ISSN system. He is in charge of the planning and implementation of various projects realized in cooperation with external partners such as EDINA for “The Keepers”. He holds a master’s degree in information science and has been working at the ISSN International Centre since 1990.

Fred Guy is the service manager for SUNCAT (Serials Union Catalogue) and project manager for the JISC funded project, The Keepers Registry. He was the Project Manager for Piloting an E-journals Preservation Registry Service (PEPRS) which was successful in establishing The Keepers Registry Beta Service. Previous jobs include being a senior lecturer at Aberystwyth University, Wales, and Director of Information and Communications Technology at the National Library of Scotland, Edinburgh.

Adam Rusbridge is Coordinator for the UK LOCKSS Alliance, a co-operative community organisation to ensure continuing access to scholarly work. He contributes to several other projects at EDINA, notably the Keepers Registry and the Pilot for Ensuring Post-Cancellation Access via Nesli2 (PECAN) project. He is a member of the TRANSFER Working Group to develop best practise guidance around journal transfer processes, as with the Enhanced Transfer Alerting Service.

¹ The ISSN IC is an intergovernmental institution governed by statutes in a convention between UNESCO and France (the host country). It coordinates the ISSN Network of national centres with the aim of introducing and operating an automated system for the registration of serials, covering the full range of recorded knowledge, through the assignment of International Standard Serial Numbers (ISSNs), <http://www.issn.org>.

1. Introduction: The Good News and the Bad

The good news is that so much of the published literature in journals, newspapers and the like can be accessed anytime, anyplace by so many.² What is bad news—or at the very least is major cause for concern—is that libraries, whether national or institutional, no longer have custody of that content: their role as stewards for future generations is being undermined.

This is especially true for scholarly and scientific journals, but can also extend to government documents and to every variety of material that libraries have mission to acquire for their patrons. This news is not truly ‘new’ but the stories are still unfolding, and remedial action is required if libraries are to continue as trusted keepers of this key part of our published documentary heritage. There is little doubt that the dynamics in the supply chain have changed: what is now purchased is a strange and uncertain amalgam of ‘rights for access’ and ‘rights to content’. There is a new type of risk to manage when content is *online remotely, not on-shelf locally*. Different approaches are being considered and acted upon: self-reliance, cooperative action, outsourcing and legal deposit.

The main focus here, and the principal purpose of this paper, is about the measures being taken to ensure that there is continuity of access and assured preservation and integrity for that published content, with special socio-technical challenges faced for serials, in e-journals and the like. This uses the Keepers Registry³ as a lens with which to highlight the work and achievements of organisations which have stepped forward to act as custodians. The emphasis is placed upon disclosure rather than audit and certification.

Our diverse literature is at risk as cultural product and a store of knowledge, both the content that has been digitized and content that is now born digital and rendered in novel ways. It is vital that we all know who is doing what, and what is not yet done. This is necessarily an international matter as the intellectual product of one nation is important to all others.

The intended outcome of the paper is discussion and agreement on appropriate strategic action, internationally, nationally and locally, to implement archival arrangement for content published electronically in serials. The first priority is that libraries prompt publishers to engage actively with the archiving agencies. The second priority is that those activities are disclosed in the Keepers Registry for all to see.

2. International and National Reports and Activity: Ten Years On

It is almost ten years since the release of the *Draft Charter on the Preservation of the Digital Heritage* at the UNESCO General Conference 32nd Session in Paris on 19 August 2003, which sought to bring about “a platform for discussions and action on information policies and the safeguarding of recorded knowledge.”⁴

The shift to the digital in the intervening ten-year period has seen a dramatic increase in the number of ISSNs assigned for electronic ‘continuing resources’ by the ISSN Network. This reflects both the growth in e-serials but also the outcome of policy action to assign appropriate identification. In April 2012, the ISSN General Assembly noted that the ISSN Register had a total of about 1.6m entries, of

² This is ‘Big Picture’ statement invites qualification related to the ‘Digital Divide’ that occurs within all nations.

³ “The Keepers Registry,” accessed 29 August, 2012, <http://thekeepers.org>.

⁴ “Draft Charter on the Preservation of the Digital Heritage,” <http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=131178>.

which 97,581 (circa 100,000) ISSNs were for online continuing resources.⁵ This includes significant coverage of the major scholarly journals, with one study⁶ reporting that 96% of Science journals were online, and as many as 86% of Arts and Humanities were also online. However, the latter statistics may be skewed towards journals that are in the English language and are indexed by Thompson Reuters, omitting significant literature that is published in other languages.

This period has seen a number of stocktaking exercises and reports. What follows is report on investigative activity in the USA and UK. There has been comparable attention in other parts of the world as the significance of this topic is global. Two reports were contemporary with the UNESCO Charter, commissioned by funding bodies in the UK and USA, the Joint Information Systems Committee (JISC)⁷ and the Andrew W. Mellon Foundation:

- *Archiving E-Journals Consultancy - Final Report by Maggie Jones October 2003*⁸
- *Archiving Electronic Journals Digital Library Federation 2003*⁹

Those reports were wide-ranging and influential, covering topics on licensing as well as preservation. The former report noted in passing that publishers never expected to undertake a preservation role, and they never did with print. Both highlighted the risks associated with digital media and formats ('digital decay' such as format obsolescence and bit rot) and with single points of failure: natural disasters (earthquake, fire and flood) and forms of human folly. The latter include criminal and political action (including hacking whereby unseen changes are made) as well as commercial events associated with the publisher and supply chain, as businesses or product lines end without transfer of legal title, actual content and assured delivery.

3. In Praise of Archiving Organisations

The *Draft Charter* in 2003 also coincided with the emergence of three types of organizations willing to act as custodians on our collective behalf. Stepping forward with digital preservation programmes are national libraries, cooperative initiatives by research libraries (typically in universities) and third-party initiatives.

The first of these, national libraries, have relied upon legislation for printed publications, requiring publishers to deposit copies of every publication in a given country into at least one library, usually a national library. (In the United Kingdom, for example, legal deposit, as the process is called, has been in existence since 1610.) Typically this enabled both current public access, through a physical visit to a national library, as well as assurance of continuity of access to the printed copy. However, legal deposit of

⁵ "Report of Activity to the General Assembly," ISSN/GA/19.2, April 2012.

⁶ Research Information Network, *E-Journals: Their Use, Value and Impact* 14 (London: Research Information Network, 2009), accessed 9 May 2012, http://www.rin.ac.uk/files/Ejournals_use_value_impact_Report_April2009.pdf, accessed 29 August 2012, <http://rinarchive.jisc-collections.ac.uk/our-work/communicating-and-disseminating-research/e-journals-their-use-value-and-impact>.

⁷ "JISC," accessed 29 August 2012, <http://jisc.ac.uk>.

⁸ Maggie Jones, *Archiving E-journals Consultancy: Final Report* (n.p.: Joint Information Systems Committee, 2003), http://www.jisc.ac.uk/uploaded_documents/ejournalsfinal.pdf.

⁹ Linda Cantara, ed., *Archiving Electronic Journals: Research Funded by the Andrew W. Mellon Foundation* (Digital Library Federation, Council on Library and Information Resources: Washington, DC., 2003), <http://old.diglib.org/preserve/ejp.htm>.

electronic publications is proving to be a complex and difficult challenge,¹⁰ even to this day, in 2012. This is despite the message being put forward by IFLA:¹¹

Legal deposit is critical for the preservation of and access to a nation's documentary heritage. Publishers and libraries work together to ensure the worldwide success of legal deposit of content, irrespective of format or technology.

The National Library of the Netherlands, Koninklijke Bibliotheek (KB), serves as a pioneering example, with the early emergence of the Depot for the Dutch Electronic Publications (DNEP) having both a national role (for the Netherlands) and with leading publishers based in the Netherlands (Elsevier and Kluwer)¹² also seen to have international significance. Project work began in the late 1990s, including leadership of the EU-funded NEDLIB (Networked European Deposit Libraries) project¹³ convened by the Standing Committee of the Conference of European National Libraries (CENL). The aim was to define the basic technological conditions for a networked European deposit library. One of the main conclusions, noting the development of the Open Archival Information System OAIS,¹⁴ was that archiving should be separated from other access services performed by national libraries, including online search facilities requiring authentication and authorisation.

In the USA there was no prospect of national electronic deposit legislation, with concern among policy makers and the libraries for the larger research universities that some other form of action was required. In response and to build upon the work of earlier studies it had funded,¹⁵ the Mellon Foundation provided development funding for initiatives that took two very different approaches to e-journal archiving: the LOCKSS project at Stanford University and the Electronic-Archiving Initiative at JSTOR.

Twelve of the e-journal archiving initiatives that were then current were reviewed in the *Metes and Bounds*¹⁶ report published in 2006. These included Portico (which had emerged from JSTOR), the LOCKSS Network and CLOCKSS (which used the LOCKSS software) as well as digital preservation programmes in the national libraries in Australia and Germany and the province/state-wide action in Ontario (Canada) and Ohio (USA). That report doubted the extent to which the Dutch experience could and should be generalized and would not replace other e-journal archiving agencies to any large extent.¹⁷

The *Metes and Bounds* report was also influential and although avowedly written from the point of view of academic libraries in North America it contains a useful list of recommendations. Some of these recommendations were directed at publishers and at the archiving agencies, others were for academic

¹⁰ Peter Burnhill and Fred Guy, "Piloting an E-journals Preservation Registry Service (PEPRS)," *The Serials Librarian* 58 (2010): 117-126, <http://www.tandfonline.com/doi/full/10.1080/03615261003622742>.

¹¹ "IFLA," <http://www.ifla.org/en/publications/ifla-statement-on-legal-deposit>.

¹² In 2002 the KB became the first official digital archive for Elsevier Science e-journals; in 2003 the KB also signed a long-term digital archiving agreement with Kluwer Academic Publishers.

¹³ Titia van der Werf-Davelaar, "Long-term preservation of electronic publications The NEDLIB project," *D-Lib Magazine* 5 (1999), <http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html>.

¹⁴ For more information, see: "Space data and information transfer systems -- Open archival information system -- Reference model ISO 14721:2003,"

http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683; and "WIKIPEDIA Open Archival Information System," http://en.wikipedia.org/wiki/Open_Archival_Information_System.

¹⁵ Archiving Electronic Journals.

¹⁶ Anne R. Kenney, Richard Entlich, Peter B. Hirtle, Nancy Y. McGovern, and Ellie L. Buckley, et al., *E-Journal Archiving Metes and Bounds: A Survey of the Landscape* (Washington, DC: Council on Library and Information Resources, 2006), <http://www.clir.org/pubs/reports/pub138/pub138.pdf>.

¹⁷ *Ibid.*, pp. 21-22.

libraries/institutions urging that they should become members of, or participate in, at least one e-journal archiving initiative.

Included in that and previous reports was a focus upon establishing the criteria for assessing what is a trusted repository of digital content, with repeated emphasis upon audit and certification. That is not discussed here—but a trail of references can be found at the Digital Curation Centre.¹⁸ There was also a proposal for a registry of archived scholarly publications that indicated which archiving agencies had preserved them, and which publications were still at risk. It is the latter and a focus upon trust through self-disclosure that is discussed below.

4. The PEPRS Project and the Keepers Registry

In 2007 JISC commissioned a scoping study on the desirability and feasibility of such a registry,¹⁹ the results of which were positive and indicated support amongst research libraries in the UK. The scoping study recommended that the purpose of the registry should be to enable librarians and policy-makers to discover the archival provision for an e-journal (which archiving scheme has been used, including access/release arrangements) and which e-journals are not (yet) within an archiving scheme. The study considered that it was feasible and recommended that this registry could be built upon SUNCAT,²⁰ the union catalogue of serials for the UK research and university libraries, considering it might be free at the point of use ('freely-available'). The evidence collected in the study indicated the importance expressed that the registry should be accurate, up-to-date and comprehensive, and that it was desirable that other serials lists and serial union catalogues should be allowed to cross-reference registry entries in order to highlight 'endangered' e-journals.

JISC acted on these recommendations. EDINA put together a proposal based upon a partnership with the ISSN International Centre (ISSN-IC), the two organisations having previously worked together in an EU-funded project²¹ and since: access to the ISSN Register was regarded as an essential component for the registry. Funding began for a project to "pilot an e-journals preservation registry service" (PEPRS) in August 2008. The aim of the PEPRS project was two-stage: to investigate and then to build and test an online facility that aggregated self-statements by organisations engaging in digital preservation of e-journal content.

A succession of demonstrator systems were created using Ruby on Rails (an open source web framework) in order to cross-match sample metadata supplied by CLOCKSS, e-Depot, Portico and UK LOCKSS Alliance against the authoritative data in the ISSN Register. This enabled some detailed work by ISSN-IC at the serial level and acted as a proof of concept for development of the user interface that searched on journal titles and retrieved information on which, if any, of the agencies included that in their preservation programme, including listings of the actual volumes of journals as held by the agency. Screenshots of the demonstrator were made available on the project website. The development proved

¹⁸ Digital Curation Centre, accessed 30 August 2012, <http://www.dcc.ac.uk/resources/tools-and-applications/trustworthy-repositories>.

¹⁹ Sue Sparks, Hugh Look, Adriane Muir, and Mark Bide, *Scoping study for a registry of electronic journals that indicates where they are archived* (Rightscom and Loughborough University, 2008), <http://www.jisc.ac.uk/media/documents/programmes/preservation/ejournalregstudy.pdf>.

²⁰ EDINA is responsible for the delivery of SUNCAT, <http://www.suncat.ac.uk>.

²¹ The CASA (Co-operative Archive of Serials and Articles) Project was funded under the Telematics for Libraries' group of projects in the EU Fourth Framework (LB-4058/B-CASA).

successful. An initial reference paper was published,²² the Abstract Data Model for which is reproduced below (Figure 1).

Abstract Data Model: Figure 1 in reference paper in *Serials*, March 2009

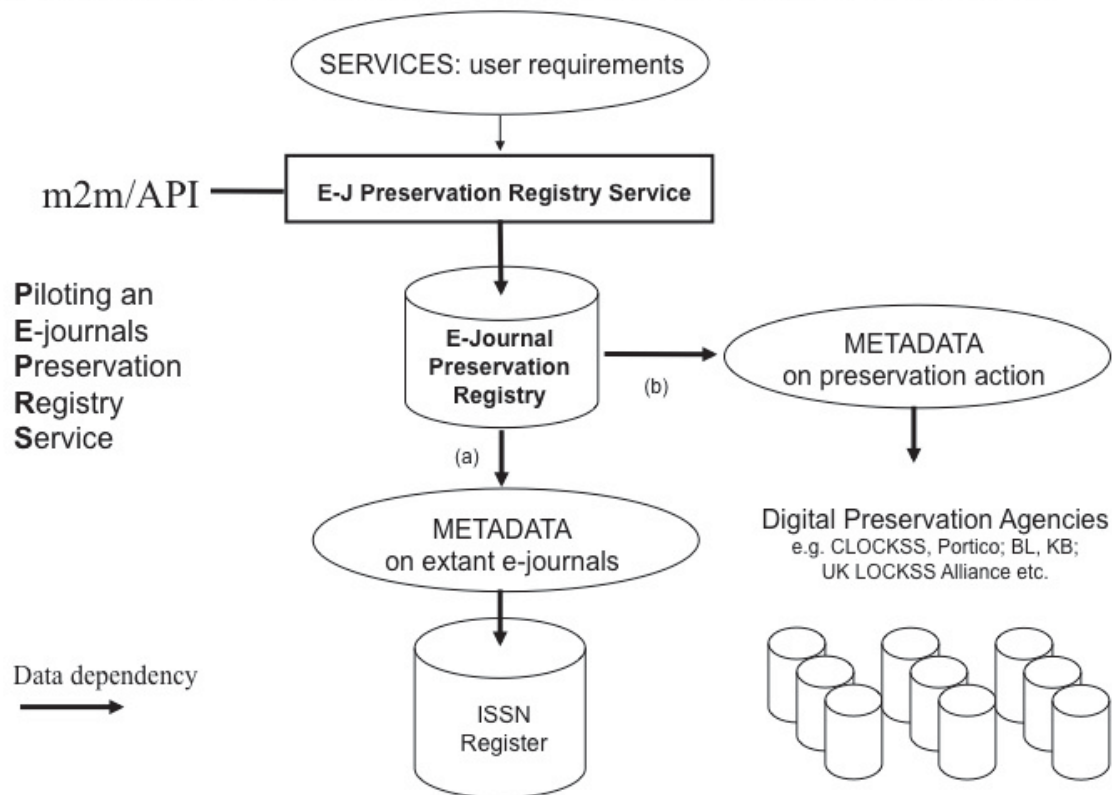


Figure 1. Abstract Data Model for PEPRS and the Keepers Registry.

Reports on progress of the project are hosted on the project website²³ as are various presentations including those made in China, Germany, UK and the USA in order to raise international awareness. The first online release was launched formally in April 2011 at the meeting of the ISSN Governing Board in Paris, as the PEPRS Beta service and generated valuable feedback from across the world. As discussed earlier, we have characterised three types of digital preservation agency: network-level organizations, national libraries, and library co-operatives. Contact had been made with organisations in each category and the following kindly assisted during the project phase:

Two that are network-level organizations:

²² Peter Burnhill, Françoise Pelle, Pierre Godefroy, Fred Guy, Morag Macgregor, Christine Rees, and Adam Rusbridge, "Piloting an e-journals preservation registry service (PEPRS)," *Serials* 22 (2009) 53-59, <http://uksg.metapress.com/link.asp?id=350487p5670h0v61>.

²³ "PEPRS," <http://edina.ac.uk/projects/peprs/>; "The Keepers Registry Blog," <http://thekeepers.blogs.edina.ac.uk>.

- CLOCKSS:²⁴ a network of archive nodes hosted in libraries around the world managed and overseen by an international Board of libraries and publishers.
- Portico:²⁵ operated and overseen by the not-for-profit organization ITHIKA and its Board of Trustees.

Three are managed as part of national libraries:

- e-Depot at KB, Netherlands
- The British Library, UK²⁶
- The National Science Library, Chinese Academy of Sciences

Two are cooperatives formed by libraries, largely in academic institutions:

- The Global LOCKSS Network:²⁷ maintained by Stanford University with funding provided by libraries that join as members of the LOCKSS Alliance.
- HathiTrust:²⁸ a partnership of academic and research institutions own and operate a collaborative digital repository, created to preserve and provide access to millions of volumes digitized from their library collections and other sources; it is administered by Indiana University and the University of Michigan.

The addition of HathiTrust was recognition that the scope of the Keepers Registry had to be wider than the e-journals from publishers and that the digitized content of print journals had to be included. This has also proved insightful for the ISSN Network, both for its decision making on the assignment rules for such digitized serial content and on planning for that assignment. There are many extra challenges of different sort with reporting on archiving of digitized journal content, including problematic volume information as well as lack of identification assignment.

As with all archiving schemes, attention also has to be given to the terms of access. And this is prompt to admit that there is still much to do on establishing agreement on the vocabulary to be used in self-statements about archival policies and practices. This is clearly where the Keepers Registry needs to gain leverage from the work of those who have been developing audit and certification schemes.

5. Enacting the vision

The recommendations for a registry made in the JISC Report in 2003 and the CLIR Report in 2006 became a reality in 2011, at least as a Beta Test Service. That was launched at the ISSN General Assembly in the UNESCO Buildings in Paris in April of that year. With a re-branding and improvement to usability and functionality following feedback, the Keepers Registry service was re-launched, still Beta mode, at the ISSN National Directors meeting in Sarajevo, Bosnia-Herzegovina in October 2011. It is live now at <http://thekeepers.org>. The screenshot in Figure 2 illustrates the way the Keepers Registry acts as a

²⁴ "CLOCKSS," <http://www.clockss.org/>. As disclosure, the University of Edinburgh, of which EDINA is part, is a founder member of CLOCKSS and acts as one of the three Archive nodes in Europe; a further three are in Asia/Pacific, one in Canada and five in the USA: none at present is in Africa, the Arab States or Latin America and the Caribbean.

²⁵ "Portico," <http://www.portico.org/>.

²⁶ "The British Library," <http://www.bl.uk/aboutus/stratpolprog/legaldep/#elec>.

²⁷ "LOCKSS," <http://www.lockss.org/>.

²⁸ "HathiTrust," <http://www.hathitrust.org/about>.

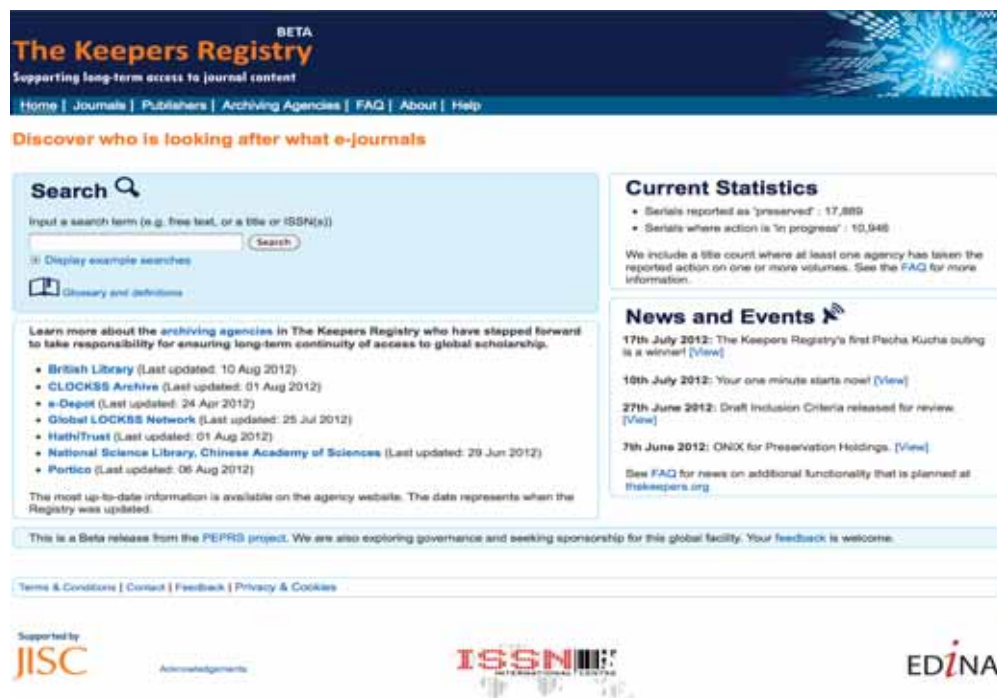


Figure 2. The Keepers Registry [screenshot]
(taken from the Beta release of the Keepers Registry on 29/08/12).

showcase for the actions being taken by digital preservation organizations, as an online and open record of their archival action on e-journal content.

It has search and browse functionality on title, ISSN and publisher. Use of the ISSN-L as the kernel identifier in the system allows search on the ISSN for the print to be entered to establish whether the electronic version of a serial title is being preserved.

The good news is that, by the time of writing (August 2012), seven of the world's leading archiving agencies report that nearly 18,000 unique Serial Titles are 'preserved'—that is one or more actual volumes—and a further 11,000 are 'in progress'.

Figure 3 illustrates the display of the volumes held at the archiving organizations for a given serial. There is some overlap in action across the archiving agencies for this particular serial, which is to be welcomed.

However, there is not good news at the volume level: the extent being preserved for any given Serial varies greatly and is far from complete.²⁹ Moreover, recall that ISSNs have been issued for 100,000 electronic continuing resources. The job is only part done.

Securing the cooperation of the archiving organisations was and is a priority: these are the organisations that do the work and they have exacting business models that cannot easily absorb extra cost. It was perhaps plain from the outset that the Keepers Registry had to be regarded as an international facility, even though it has been funded by JISC for the UK, as had been predicted in the scoping study:

²⁹ A surprising welcome side-benefit has been contribution to the development of new standards ONIX for Preservation Holdings (ONIX-PH), <http://www.editeur.org/127/ONIX-PH/>, and the Universal Holdings Statement.

The Keepers Registry BETA
 link to The Keepers Registry search page
 Supporting long-term access to journal content

Home | Journals | Publishers | Archiving Agencies | FAQ | About | Help

Search > Search results

1532-2416 Search

1 hit found
 Current search: 1532-2416 (Key Words)

Title	ISSN	Publisher	Current extent of archiving	Archiving agency
1. Communications in soil science and plant analysis	1532-2416 (Online); 0010-3624 (Print)	Taylor & Francis	Summary of extent not yet available	CLOCKSS Archive
		Taylor & Francis	Preserved: v. 1-39	e-Depot
		Taylor & Francis	In progress: v. 1-41, 43-present	Global LOCKSS Network
		Taylor & Francis	Preserved: v. 42	Global LOCKSS Network
		Taylor & Francis Group	Preserved: v. 28-42	Portico

Page 1 of 1

Figure 3. Archived content as reported to Keepers Registry [screenshot]
 (taken from the Beta release of the Keepers Registry on 29/08/12).

It seems to us that in order to gain the co-operation of the archiving organizations based around the world, which would be vital to its utility, the registry/registry-like service would have to be conceived as something which would serve the whole international scholarly community.³⁰

The inclusion of CLOCKSS, e-Depot, LOCKSS and Portico, all of which originate outside the UK, as well as the British Library was both deliberate and essential. This also motivated the early inclusion of the National Science Library of China (NSLC) with which EDINA had previous working contact and exchange of staff. Waiting to join the Keepers Registry are a cooperative organization in Canada, a large-scale university library in the USA and a discipline-specific data organization.

The Keepers Registry seems on course to become a global monitor of international content of significance for each and every country, as well as a showcase for the archiving agencies. Indeed, the Keepers Registry may be that key component in the digital infrastructure that Donald Waters was looking for in 2002 when he wrote:

Our vision is much less clear about the infrastructure needed to enable archives to cooperate and interoperate.³¹

The good news is that participation in the Keepers Registry prompted discussions amongst the archiving organisations, with calls for greater social media functionality in order that the Registry be function as a 'safe places network' enabling exchange between those organisations with archival intent, and also assisting engagement with others who care about the issues involved: not only the titles and extent of volumes actually preserved, but also the variants in technical approach, business model and the terms of availability for triggered content.

³⁰ Sparks et al., "Scoping study for a registry of electronic journals," p. 32.

³¹ Donald Waters, "Good Archives Make Good Scholars: Reflections on Recent Steps Toward the Archiving of Digital Information," in *The State of Digital Preservation: An International Perspective* (Washington, D.C.: Council on Library and Information Resources, 2002), 78-95, <http://www.clir.org/pubs/abstract/pub107abst.html>.

In such a conference organised by both UNESCO and IFLA, there should be good advice on hand about how really to be an international facility. There is experience gained via the ISSN Network in dealing with national libraries but we have considered it important to engage with research libraries more generally. There has already been outreach to research libraries through LIBER (for Europe) and ARL (for Canada and the USA). Is there sense in approaching this ‘regionally’, building on the approach made with respect to other regions? The approach taken may assist thinking about governance and the long-run business model.

Certainly the use of the ISSN Register within the Keepers Registry is an essential part of this international role as well as what is required for any national role. The connection to the members of the ISSN Network, mostly hosted in national libraries, is also valuable.

The intention to be a global facility has renewed focus upon global usability, and the language in the user interface for the Keepers Registry beyond that of plain English in order to improve the results of automatic translation software, seeking semantic equivalents of Keeper, stewardship and custodian. Feedback on that and offers of help would be welcomed: assistance from UNESCO Volunteers has been mentioned.

6. Forward Look

The literature required and consulted in one country will often have been published by another. This prompts return to the doubts raised in the *Metes and Bounds* report³² on the sufficiency of legal deposit and noting following concerns:

1. Differential ability of national libraries to take on digital preservation as part of legal deposit
2. Complexity where a company in one country has an editorial board (and readership) in another, and its servers in a third
3. The resultant haphazard mix of compulsory, voluntary and lack of schemes across the globe
4. Physical access restrictions (having to visit the particular national library).³³

It is evident that research libraries in the universities have had to act, and not wait upon national libraries. This is true for both for e-journal content that is born digital and for print journals that have been digitized, with obligation to advertise systematically the latter—a role that the Keepers Registry is playing for the content in HathiTrust.

About 20% of the ISSNs issued by members of the ISSN Network for electronic continuing resources have been issued by the Library of Congress with respect to place of publication in the USA. The British Library has assigned about 10% of the total for those published in the UK. Netherlands have each assigned about 4.5%. The archiving organizations being monitored by the Keepers Registry are based in these countries, which in aggregate cover over a third of those digital resources, which have an ISSN. Of course those organizations engage with publishers in other countries. However, there is surely shared priority in university, research and national libraries working together to ensure that publishers in regions across the world engage with the existing archiving organizations. And prompt those

³² Kenney et al., *Metes and Bounds*, pp. 21-22.

³³ The entry in the Keepers Registry for e-Depot: “Licensed content can only be accessed onsite at the KB. Open access and triggered content is freely available via the online e-Depot portal.”

organizations, including national libraries that embark upon preservation programmes, to report their activity through the Keepers Registry.

There is another compelling reason why there is no time to wait upon the law to catch up. The focus here has deliberately been upon monitoring archival actions on what is now ‘traditional digital’: scholarly and scientific journals, government documents and the like. However, as the Web becomes the principal arena and medium for scholarly discourse, the problem space has become much richer, as well as more challenging. There is much that is issued on the Web but nowhere else. Scholarly statement and government report contain data and multimedia. This is especially important when considering what is the copy of record.

What is on the Web is now referenced and cited in support of scholarly analysis and statement in e-journals. The Web is dynamic: what is at the end of a given Uniform Resource Locator (URL) can and does change. This topic, termed ‘citation rot’, is being investigated by a team led by Herbert Van de Sompel (Los Alamos National Laboratory) who provides overview in a recorded talk given to the STM Innovations Seminar 2011.³⁴ It is important to know whether what was referenced has ceased to exist or has been archived somewhere. Using Memento,³⁵ which enables ‘travel back in time’ by searching the Internet Archive, a recent study³⁶ reviewed articles in two scholarly repositories in order to establish which cited resources were still current and which were not, and whether the content that existed at the time the citation was made had been archived. “The most dramatic finding is that 45% (66,096) of the URLs that currently exist [in one subject repository] are not archived at all [and 28% of the resources referenced by the articles in the institutional repository had been lost].

This is a challenge, not only for the Keepers Registry.³⁷

³⁴ Herbert Van de Sompel, “Time Travel for the Scholarly Web” (talk given at STM Innovations Seminar 2011, Hilton London Kensington Hotel, London, UK, 2nd December 2011), <http://www.stm-assoc.org/events/stm-innovations-seminar-2011/>, recorded at <http://river-valley.tv/time-travel-for-the-scholarly-web/>.

³⁵ “Memento,” <http://www.mementoweb.org/about/>.

³⁶ Robert Sanderson, Mark Phillips, and Herbert Van de Sompel, “Analyzing the Persistence of Referenced Web Resources with Memento,” submitted to *arXiv* on 17 May 2011, arXiv:1105.3459v1.

³⁷ Peter Burnhill, “Tales from The Keepers Registry: Serial Issues About Archiving & The Web,” *Serials Review* (forthcoming); The University of Edinburgh, including EDINA, have been successful with a funding proposal to work with the Memento Team at Los Alamos National Laboratory to carry out a large-scale investigation with recommendations for archiving cited Web content.

The World Audiovisual Memory: Practical Challenges, Theoretical Solutions?

Treasures That Sleep

Film Archives in the Digital Era

Jean Gagnon

Director of Collections, Cinémathèque québécoise, Montreal

Abstract

With the all digital approach in audio visual preservation, and with the disappearance of film labs and film stocks, it is the very materiality of film that is vanishing. But if we want to preserve access to these films, if we want to restore them with a consideration for their authenticity, understanding of this materiality, photo-chemical processes, colour processes, special effects technics, etc., will need to be preserved.

Author

Jean Gagnon has a passionate interest in books and archives and over 20 years' experience in audiovisual collection and archives management. After earning a bachelor of fine arts degree, with a major in film production and film studies, he worked for three years at the Canada Council for the Arts and spent the next seven years as associate curator of media arts at the National Gallery of Canada, followed by ten years as executive director of the Daniel Langlois Foundation for Art, Science, and Technology. He has also taught at Canadian universities and served as a consultant to cultural organizations. Besides being the Director of Collections at the Cinémathèque québécoise, he is working toward a doctorate in art studies and practice at l'Université du Québec à Montréal.

I have been asked to be the “film preservation” person here, but in my daily work at the Cinémathèque québécoise with the collections I’m responsible for I have to deal with film-based material, video or magnetic elements as well as digital files, plus non-film collections (scripts, photographs, paper archives and so on). But to play my role, I’ll start by pointing a great irony. At the very moment when film-based audiovisual production, post-production and distribution is definitely receding, if not totally disappearing, film-based material is still considered by many as the best storage media, not only for moving images, but also for data. Apparently, the US Army stores its most precious data on film. Major Hollywood film producers now distribute their films as Digital Cinema Packages (DCP), but they still preserve them on film, actually on three black and white film with colour separation which is very good but expensive. Film lasts more than a hundred years and it will be relatively stable if stored in appropriate conditions,¹ while digital files are fragile and dependant on numerous complex and, to a certain extent, unknown factors such as technology obsolescence, material and magnetic decay, operating systems and software/hardware changes and evolutions, bit loss, and so on. Accessing these files in the future will cost lots of efforts over, say a 100 years, and will require a lot of financial investments in technologies that are not yet invented. Another irony, is that the stable medium of film is now among the endangered species of the audiovisual kingdom because, soon enough, projectors, expertise in film projection and handling, not to mention knowledge about the photochemical aspects of film and film printing will disappear, rendering difficult access to the content in its original form. Here it is worth recalling the introduction to the FIAF’s Code of ethics:

¹ The Image Permanence Institute in Rochester (NY) estimates that colour film and negatives stored at -5 Celsius, 30% relative humidity, could be preserved and suitable for viewing for 2,474 years.

Film archives and film archivists are the guardians of the world's moving image heritage. [...] Film archives owe a duty of respect to the original materials in their care for as long as those materials remain viable. [...] Film archives recognise that their primary commitment is to preserve the materials in their care, and - provided always that such activity will not compromise this commitment - to make them permanently available for research, study and public screening.

Not only cinémathèques have to preserve film reels, they now have the added duty of preserving the film experience as a projection, as the performance of a work relying on an obsolete technology.

Anyhow, what is referred to as the Digital Era generates a lot of anxieties and questions among film archivists when it comes to the fate of cinema and film heritage in the future. The cinema era span more than a hundred years and not only it is a major means of artistic expressions and a major cultural force shaping and influencing contemporary global cultures, but also it is the very memory of most of the 20th Century documenting how peoples, societies and cultures evolved, moved, fought, celebrated, danced or sang.

This memory on film is dispersed and not all is already under the custody of an archive or other responsible body. And what is in the care of such custodians is hard to estimate in terms of hours or number of titles. [SLIDE] This can be explained in part by how cinémathèques have collected or how material entered their holdings: from labs during their time of operations or when they closed, from producers and distributors, from governments and other official bodies, from individuals, be they experimental filmmakers, artists, or you and I with our family films. Often these entered collections in great quantity, all at once, and were not catalogued at the item level for lack of personnel to do so. It is therefore very hard to know world-wide what is at stake. But we can suspect that quantities would be staggering. A study² published in 2010 by *Screen Digest* estimates that 42.7 millions hours rest in world archives,³ while the largest of these only digitize 1% of their holdings annually. As an example, the Cinémathèque québécoise has more than 115,000 records in its database of international film, television, and video corresponding to approximately 65,000 titles. And 1% would be a fair estimate of our rate of digitization at this point.

In Canada right now there is basically no money to digitize films unless a producer wants to repackage his films for a DVD or blue-ray release or for pay tv. The National film board of Canada is a good example of a producer taking care of its film assets. It is in the process of digitizing all its catalogue and planning the storage of its original, film-based material. They apply very high standards of digitization and storage (2K for 16mm and 4k for 35mm) and after this process the film material returns to storage. I insist here on storage of the film-based material, because, film being a more stable technology, it is important to store the original properly and catalogue it precisely. Storage in controlled atmosphere as a costs (like energy costs), but in comparison with digital preservation it will be more cost efficient for the foreseeable future.

Another important digitization project in Canada is a private enterprise called Elephant, memory of Québec Cinema. Based in Montreal, it is a small (dollar-wise) cultural investment by Quebecor, the media giant. Their aim, as explained on the Web site: "Within Elephant: memory of Quebec cinema, all Quebec

² Claire Harvey, "The Global Trade in Audio-visual Archives," *Screen Digest*, 2010, 40 pp., http://www.screendigest.com/reports/201074c/10_08_the_global_trade_in_audio_visual_archives/view.html.

³ Although it not clear what percentage of this number is constituted of film versus video elements.

feature films will be gradually carried on to digital media (in both standard and HD [definitions]).”⁴ They digitize and restore feature length fiction films in order for them to become available for the public on the Illico pay-per-view channel of Videotron, one of Quebecor’s companies. The reality of such an endeavour is that, although all film elements used to do this work come from the Cinémathèque québécoise, we have little involvement in the actual process employed to digitize, restore or keep the final restored versions, which are not deposited with us. None of the restored films in this project is the object of a return to film-based material for their conservation and for their screening. Thus, we have no precise data about methods, standards applied and so on from Elephant. Given their final output which is HD television certain things such as colour and contrast calibrations may not be the same as for a theatre release. Elephant’s activities are commendable as it is one of the rare restoration project in the country going on right now, but it is done rather as a private endeavor with no or very little transmission of knowledge and data to the archival community.

There is other reasons to preserve and document knowledge and expertise about film-based material and technologies over the history of cinema. Scholars of film have started to be much more attentive to film technologies, including the different systems of colour film stocks and their processing. Concerning peculiarities of film-stock, Luca Giuliani and Sabrina Negri mention the case of the restoration of Jacques Tati’s *Jour de fête*. They write:

The film was shot in 1949 with two adjacent cameras, one which was recording on standard black and white stock, and another set up to shoot on a color stock called Thomson color. Thomson color was a lenticular color system, based on the same patent as Kodak’s 1928 Kodakolor. The functioning of the lenticular color system was very complicated.⁵

I stop the quote there, but the authors go into longer details about a complex system of colour deconstruction (at the shooting stage) and colour reconstruction (at the projection stage). I want here to point out that the preservation of film-based material and the consideration of its material base, in this case colour systems, documenting these aspects is important when it comes to restoration work with or without a return to film. So again preservation of knowledge and expertise, transmission of those through education is of utmost importance.

Alexander Howarth, the Director of the Austrian Film Museum, said quite provocatively at a conference in 2011 at the Cinematheque française:⁶

...from now on – instead of spending billions for digitization projects that primarily serve the industry – public cultural budgets and political energies need to be activated in order to ensure the continued production of film stocks and printing and projection machinery as well as the perpetuation of all related professions and systems of training. All this mainly for museum and heritage purposes, so that film, like many other historically

⁴ http://elephant.canoe.ca/a_propos/, only in French.

⁵ Luca Giuliani and Sabrina Negri, “Missing Links: Digital Cinema, Analogical Archives, Film Historiography,” *Intermedialités* 18 (Fall 2011): 71-84. (Paper presented at the international conference titled “The Impact of Technological Innovations on Historiography and Theory of Cinema,” Montreal, Quebec, November 1-6, 2011)

⁶ October 14th, 2011, at the Symposium entitled *Révolution numérique: et si le cinéma perdait la mémoire*. Published as Alexander Howarth, “Persistence and Mimicry: The Digital Era and Film Collections,” *Open Forum* 86 (April, 2012): 21-27.

influential art practices, can be preserved and kept visible as such and not only as a digital version of itself.⁷

At the present moment one of the last remaining film lab in Canada is that of the company Vision globale in Montreal. It is the only lab on the East-coast of North-America and one of the few left on the whole of the continent. The company tells us it might close the lab in two or three years from now! Can we imagine saving a lab? And the expertise that goes with it? Is it too farfetched? It seems that it was not for the Swedish Film Institute that did just that; it acquired the last lab in Scandinavia and incorporated some of the key staff with the aim of transmission of skills, and acquired the technology. But I think one of the key elements of the SFI's scheme is the concurrent establishment of a lab for digital cinema so that the two worlds cross paths.

This shows that instead of seeing this insistence on the conservation of film and the documentation and transmission of expertise and technologies of the film era as a backward endeavour, while the present and the glowing future of cinema is already digital; that instead of seeing these as opposite sides of a battle, we can bridge the gap if we see them for what they provide : film preservation, taken seriously, is conserving and safeguarding original material and the capacity to perform the works according to their technical specificities and intended aesthetics, and digital technologies are tools for restoration and increased access.

I hope that I was able to open the frame of reference in this discussion which is often one sided; everything can be digitized, based on the distinction between the carrier and the content, with very little consideration for the material and physical base of what it is we preserve. Here, the carrier, film, might necessitate due consideration and I would greatly advise that a priority is still to invest in proper storage of film-based material and not to rely solely on digital means to save film heritage.

⁷ Ibid., p. 27.

Seeing, Hearing, and Moving Heritage

Issues and Implications for the World's Audiovisual Memory in the Digital Age

Caroline Frick

Abstract

Recorded sound and moving images have captured some of the most influential human achievements, daily events, and tragedies of our world. Although advocates for universal access to global heritage receive public accolades, the audiovisual archives community, in charge of millions of artefacts in need of reformatting for simple playback, remains acutely aware of key questions: Who will pay for such access, and how can we ensure long-term support for such projects and long term preservation? This paper will provide an overview of the current challenges faced by the world's audiovisual archives, in particular that of raising public awareness, as well as announce a new global initiative encouraging cooperation and collaboration for the sake of audiovisual digitization and preservation.

Author

Caroline Frick is an Assistant Professor in the Radio-TV-Film Department at The University of Texas at Austin and is the founder and Executive Director of the Texas Archive of the Moving Image: www.texasarchive.org in the United States. Dr. Frick worked in film preservation at Warner Bros., the Library of Congress, and the National Archives in Washington, D.C. Dr. Frick also programmed films for the American Movie Classics cable channel in New York and currently serves as the President of the Board for the Association of Moving Image Archivists. Her book, *Saving Cinema*, was published in 2011 by Oxford University Press.

1. Introduction

Since the inception of moving image and recorded sound technologies in the late nineteenth century, audiovisual materials have assumed a central role in human culture around the world. Our primary understanding of global events is more comprehensible and enriched via moving images and sound. An inspiring, eclectic range of audiovisual collections—from Hindi language feature films that appeal to global audiences, audio recordings of languages now gone, to filmed and recorded testimonies of genocide survivors—often share a similar fate: decomposition, destruction and disappearance from our world's collective memory.

In the United States alone, over 50% of Hollywood-produced moving images created before 1950 no longer exist.¹ While shocking, that particular Library of Congress-produced statistic only references Los Angeles corporate-produced product. When scholars and archivists team up to look more seriously at the breadth of American “movies,” from industrial and educational films, documentaries, home videos, etc..., statistics reveal even more significant losses that are virtually incalculable. At the World Electronic Media Forum which took place in Tunis in late 2005, UNESCO noted that approximately 80% of the 200 million hours of footage known to be in the world's television archives will likely disappear by 2015.² Data

¹ “Film Preservation 1993: A Study of the Current State of American Film Preservation: Report of the Librarian of Congress,” (Washington, D.C.: National Film Preservation Board of the Library of Congress, 1993), p. 3, <http://www.loc.gov/film/study.html>.

² Sue Malden, “Archives@Risk” (presentation, Georgetown University, August 2007).

compiled by the United States of America's Institute of Museum and Library Services (IMLS) indicate that in that country's public institutions, nearly 17,000 organizations held sound recordings "at risk," nearly 50% of which remained in "unknown condition."³ These are statistics that reflect merely the material thus known in archives; the vast majority of the world's audiovisual material remains uncollected and far outside conventional repositories or professional expertise. **For audiovisual preservation, the clock is ticking.**⁴

2. Digitization and Preservation Obstacles:

A number of important factors contribute to the uniquely vulnerable state of both analogue and digital audiovisual material. Two other papers presented at "The Memory of the World in the Digital Age: Digitization and Preservation Conference" will be looking more closely at the specific technical challenges to sustaining film, television and digital A.V. files. My goal here is to introduce very broadly the variety of topics affecting the preservation of audiovisual materials as well as to highlight one too often forgotten obstacle: the need to increase public awareness and support for preservation efforts.

The ever-quickenning obsolescence of playback equipment threatens decades-worth of historical, aesthetic and cultural heritage in communities in the developing world as well as the most affluent European cities. The disappearance of even the most seemingly ubiquitous playback equipment alongside the cessation of parts manufacturing, combine to threaten the current preservation efforts of archivists, particularly with electronic analogue formats. Without the ability to play back electronic media, there can be no digitization, no preservation and no access; millions of hours of televised and filmed events and activities from every corner of our world will be lost forever.

In September of 2012, the Fuji Corporation announced that it would cease production of nearly all of its motion picture film, a medium still considered by the global film preservation community to be the ideal long term preservation solution for the moving image. Although percentages vary depending on the source, Fuji provides between 20% and 40% of the world's motion picture stock, Eastman Kodak's promise to step in to supply even more than the film they currently manufacture seems shaky at best with the company's own bankruptcy filing this same year. Many filmmakers and media executives, however, watch without concern as celluloid, a beautiful but stubbornly 19th century technology slowly fades away during the 21st century, because of the omnipresence and seeming ease of digital capabilities.

Unfortunately, those who glibly advocate for a quick and easy "digital transition" for moving images would heed well the warnings detailed in a 2008 joint report on digital preservation between repositories and research organizations in the United States, United Kingdom, Australia and the Netherlands: "Embodying creative works in digital form has the unfortunate effect of potentially decreasing their useable lifespan. Digital information is ephemeral: it is easily deleted, written over or corrupted. Digitized and born digital materials are an important part of the world's cultural heritage, but unless active steps are taken to preserve them, they will be lost."⁵ The persistent lack of a concrete

³ Rob Bamberger and Sam Brylawski, "The State of Recorded Sound Preservation in the United States: A National Legacy at Risk in the Digital Age," report commissioned by the National Recording Preservation Board of the Library of Congress (Washington, D.C.: Council on Library and Information Sciences and the Library of Congress, August 2010), 10, <http://www.loc.gov/rr/record/nrpb/pub148.pdf>.

⁴ UNESCO World Day for Audiovisual Heritage 2012 theme.

⁵ "International Study on the Impact of Copyright Law on Digital Preservation," a joint report of the Library of Congress National Digital Information Infrastructure and Preservation Program (US), the Joint Information Systems Committee (UK), the Open Access to Knowledge Law Project (AU), the SURF foundation (NL), July 2008.



Figure 1. Images of women in Thailand working to clean obsolete video formats.

preservation standard for digital audiovisual material, particularly moving images, presents only one of the many factors related to the reality of digital technology's ephemeral nature.

Technology, in its most basic sense, remains a central, and certainly the most easily understood challenge for both corporate and non-profit custodians of the world's audiovisual heritage. But it is critical to note that such technological obstacles merely serve as the first of even more pressing constraints. Regardless of the technology at the core of a project (e.g., a 35mm Steenbeck, a Umatic video deck, or a DVD player from the 1990s), the strains on human labor, both financial and emotional, required for digitization and preservation associated with such tedious work likely remain the most pressing, most expensive (and definitely most difficult to fund) cost. Workers in the smallest of entrepreneurial audiovisual libraries and archives to technicians in massive digital asset management companies share the strain of repetitive, detailed and physically draining work. Many audiovisual artefacts require "real time" transfers which can be additionally problematic to workflow, financial costs, and simple staff or personnel morale.

The economic impact of digitizing and hosting access portals of audiovisual becomes increasingly complicated as most archives and archivists are unable to digitize in-house and must work with boutique, expensive vendors. When governments do prioritize the digitization of audiovisual material, little attention is given to the sustainability (aka, long term preservation) of such projects. Legal constraints of corporate-produced material (specifically copyright protected music and film) prevent adequate access to preserved and already digitized media in major non-profit or government repositories. Perhaps even more significantly, audiovisual artefacts that were registered for copyright by specific owners who can no longer be found or contacted, remain "orphaned" and virtually unusable in repositories and community collections all over the world.

In essence, the challenges facing the global audiovisual preservation community closely reflect, even mirror, the challenges laid out for attendees and participants by the convenors of the "Memory of the World in the Digital Age: Digitization and Preservation Conference." It is critical to highlight a particularly important aspect underlying this long list of obstacles, and, frankly, one too often ignored by

archival practitioners: The need to increase awareness of the problem itself to the world's constituencies who make and consume audiovisual material at an accelerating rate, thanks to increasingly affordable media-making technologies. Outside of rarified practitioner circles, challenges to the preservation of film, broadcasting or recorded sound collections (much less the stark statistics of the significant global losses already) are met with confusion, not concern.

Arguments for the preservation of ubiquitous, contemporary and seemingly “glamorous” audiovisual media provides a stark contrast to expressing passion for preserving Lebanon's Phoenician Alphabet, the seventeenth century baptism records of slaves from the Dominican Republic, or medieval tapestries of Europe.⁶ How, people might wonder, does *The Wizard of Oz*, produced in 1939 by MGM Studios in Hollywood, CA and seen the world over in multiple variations, digital, celluloid or otherwise, merit preservation attention or, indeed, its inclusion on UNESCO's Memory of the World Register?⁷ The answer is both complicated and simple at the same time: Although a corporate product currently protected by a corporate owner, *The Wizard of Oz* remains a beloved piece of cultural heritage in many parts of the world. Moreover, such a famous movie can be a unique “hook” to engaging a public familiar with such feature film “classics” with more esoteric, historical and fascinating audiovisual records.

When offered a public platform to discuss their avocation, audiovisual preservation professionals understandably turn first to the technical challenges of audiovisual preservation versus that of text preservation. Although the long term conservation of paper in the right archival conditions is well documented, the long term preservation of analogue and digital video is wholly unproven and offers an enormous challenge with over hundreds of thousands of hours of material captured on enormous files. What is often lost in this practitioner tendency to focus on the technical challenges, however, is the opportunity to engage the broader public, funding agencies and the media industries themselves by a focus upon the access of vulnerable content, regardless of original medium, to explain more clearly the artefact's value itself. In essence, to challenge the global AV preservation community to say “it's ok” to prioritize some access first in order to drive preservation—a direct flip to decades old professional mantras.

One key “archive at risk” is currently based in a former medical school morgue: the Universidad de San Carlos de Guatemala's Film Library, the Cinemateca Universitaria Enrique Torres (Cinemateca). At the Cinemateca, a few part time staff works to protect over 5,000 reels of film on a variety of formats—from 35mm nitrate to 8mm home movies. Founded by a member of faculty in the late 1960s or early 70s, the Cinemateca's current aims reflect its earliest iteration as a cine-club—showcasing rare and even banned films amid political turbulence. With a contemporary mission to share and to protect the country's film material and to support the growth of the nation's film production, the Cinemateca battles economic adversity as well as its climate (Figure 2).

The Cinemateca staff, when interviewed, voiced an important dissenting voice to current heritage preservation practice, arguing that they believed it vital to digitize much of their material, even with the knowledge that the digital format might be less ideal over the long term. Some kind of digitization, they felt, could drive more interest in raising funds towards longer term preservation efforts.⁸

⁶ All of these artefacts are currently on the Memory of the World register.

⁷ Nominated by George Eastman House International Museum of Photography and Film in Rochester, New York (USA), *The Wizard of Oz* was entered into the Memory of the World Register in 2007.

⁸ For more information regarding the Cinemateca's challenges and approach, see Caroline Frick's *Saving Cinema: The Politics of Preservation* (New York: Oxford University Press, 2011).



Figure 2. Images of rotting films in Guatemala's Cinemateca Universitaria Enrique Torres.

In the second decade of the 21st century, it is time to try out some new approaches to raising awareness about what so many do not even understand as a problem. The CCAAA is today launching a program intended to serve as a catalyst for action, a high profile Public Relations campaign for audiovisual associations around the world: Archives@Risk.

3. A Concrete Step Forward: Archives @ Risk

Officially formed in 2000, the Co-ordinating Council of Audiovisual Archives Associations (CCAAA) represents the interests of an array of organizations focused upon preserving materials such as broadcast television and radio, film, and audio recordings. Members include the Association of Moving Image Archivists (AMIA), the International Federation of Television Archives (FIAT), the Southeast Asia-Pacific Audiovisual Archive Association (SEAPAVAA), the International Association of Sound and Audiovisual Archives (IASA), the Association for Recorded Sound Collections (ARSC), the Federation of Commercial Audiovisual Archives (FOCAL) as well as the audiovisual interest sector of the International Council on Archives (ICA) and the International Federation of Library Associations (IFLA). CCAAA members include professionals working in for-profit and non-profit organizations, higher education, government agencies and branches as well as devoted individuals and hobbyists. CCAAA provides a communication forum as well as a platform for joint initiatives in this specialized area of heritage preservation.

The official discourse underlying the value and rationale for the creation of CCAAA accurately denotes that “the preservation of the audiovisual heritage concerns those who produce, those who preserve and disseminate as well as those who use audiovisual documents.”⁹ The CCAAA membership, embracing the world’s most populous and arguably most active A.V. preservation associations clearly substantiate this claim. Where CCAAA has struggled most in its twelve years of existence is encapsulated in a core component of its goals for action: the need to co-operate amongst these associations which represent a diverse organizational membership to offer a united program of action and priorities. In addition, with virtually every representative to CCAAA a volunteer for their own A.V. preservation organization in an era of institutional cutbacks, time is at a premium.

Understanding CCAAA’s difficult work over the last several years must be seen within the context of an often under-discussed aspect to archival work which can complicate the repeated request by funders and government agencies to collaborate: tensions and underlying competition within the preservation community itself. Archivists, librarians and museologists from any part of the world can be (and in many cases *should be* so as to protect the collections under their care) very proprietary in nature. Professional wariness over collaborative projects that are often added to daily administrative duties, even more of an acute concern in an era of limited resources where an understaffed labor force is doing more with less, can, in the worst case scenario, result in problematic and often failed joint ventures. Collaborative efforts struggle to maintain projects or cultivate sustainable support from government agencies, foundations or other types of funding bodies.

With a profound and sober understanding of such challenges, CCAAA’s annual meeting in 2012 garnered renewed support for a specific, concrete initiative that could establish preliminary steps with obtainable, escalating outcomes in the area of preserving and digitizing moving image and sound heritage collections. CCAAA offers a partnership prototype amongst various stakeholders whose shared passion and efforts can create substantive change. Thus, on behalf of CCAAA, I am pleased to announce our “Archives@Risk” program.

Archives@Risk aspires to unite a wide range of collections, archives and professionals and, in doing so, offers concrete models to challenge standard protocols of archival practice and action. CCAAA, as an “umbrella” organization between audiovisual associations, will showcase current projects from the global membership as well as launch new initiatives. Archives@Risk will bring together a variety of countries, archives, collections, and industries that might find it challenging to collaborate via traditional inter-government programs. For example, one of the key obstacles in digitization and preservation for endangered audiovisual heritage in all regions of the world is that so much valuable material remains outside of established archives, museums or libraries. Archives@Risk provides an opportunity for professional exchange in a more informal mode.

The first iteration of the Archives@Risk program was an outgrowth of exchange at the World Electronic Media Forum that took place in Tunis during late 2005. At the Forum, an ad-hoc group, including representatives from the United Nations, UNESCO, the World Broadcasting Union (WBU), the European Broadcasting Union, and the International Federation of Television Archives, formed and conducted a global survey on endangered A.V. archives and collections predicated on the WBU and their regional membership. The results, while not wholly inclusive, provide concrete data about the enormous

⁹ Co-ordinating Council of Audiovisual Archives Associations, accessed September 1, 2012, <http://ccaaa.org/what.html>.



Figure 3. Maps representing the most critical areas of the world for digitization and preservation of audiovisual materials according to a 2007 World Broadcasting Union survey.

problem at hand, and of the most endangered regions, particularly material held in West Africa and Southeast Asia.¹⁰

The initial objectives of the Archives@Risk project were ambitious. Leading organizations had been brought together to articulate more clearly both the breadth of the challenges facing the global audiovisual heritage as well as to identify key areas of vulnerability. In addition, the program aspired to create an online library of audiovisual preservation resources; to provide access to highlighted moving image and sound clips to illustrate the value of such material; and, perhaps most importantly, to save particularly endangered collections. The initial work accomplished on the project indicated great need and interest and resulted in a rich website with clips and resources. Quick action was taken to rescue two collections “at risk” in Madagascar and Vietnam. However, the original group charged with the project realized that to achieve more substantive impact, and in order to be more fully inclusive, more partners in the project were needed. Thus, the International Federation of Television Archives brought Archives@Risk to the CCAAA who now will spearhead project development and strategy.

In light of the goals for “The Memory of the World in the Digital Age: Digitization and Preservation Conference,” Archives@Risk merits attention as a model for cooperative action across industry and academia; government and non-government organizations; professionals and amateurs. Lessons learned from the initial work on the project indicate that for more substantive long-term success, Archives@Risk will need to function on a two tier level: First, the program acts as a united front for the world’s largest A.V. archive professionals to raise public awareness of the preservation challenges in the digital age as well as reinforce the value of individual collections, A.V. repositories and those that work in them. Second, Archives@Risk strives to obtain funding and to create the infrastructure for expertise, exchange and cooperation between different organizations and individuals predicated on specific A.V. digitization and preservation needs. For example, Archives@Risk will be developing further its “Archive Buddy” program that successfully partnered broadcasting giant NHK in Japan to come to the aid of a decomposing video collection in Vietnam.

CCAAA believes that the digital archives and preservation community must not forget that a critical part of the work we need to do is to continue to explain to the global public why film, television, recorded sound and borne digital versions of all such A.V. materials are critically important to our

¹⁰ Malden, “Archives@Risk.”

memory of the world and, as such, must be preserved. Archives@Risk provides the inter-association organization with a powerful public tool with which to do just that.

4. Conclusion

Audiovisual material occupies an important place in UNESCO's Memory of the World program. Currently, the MOW Register contains over fifty film and videos, from Cuba and Chile to Lithuania, the Philippines and Luxembourg, representing the eclectic nature of audiovisual heritage. Audiovisual material is more than just "the movies" or "music" in the same way that challenges to its preservation and digitization are far more than purely technological in nature.

One particularly important "non-technical" challenge shared by all involved in the preservation and digitization of A.V. archives whether a small, entrepreneurial collector or a large government funded repository, is the necessity to increase awareness and to cultivate a true sense of urgency to the plight of moving image and sound cultural heritage. We need to balance more pragmatically the needs of preservation and access, particularly in light of the critical dearth of public awareness of preservation challenges. CCAAA's Archives@Risk program will champion broader access to materials to convey the urgency of long term digitization and preservation requirements.

In a sense, A.V. archives need to learn what UNESCO has succeeded at so beautifully with its Memory of the World program: to excite the public by a broad, universally appealing awareness campaign with which to inspire and engage both today's as well as future generations. With Archives@Risk, CCAAA, together in partnership with UNESCO, aspires to raise awareness, funds and concretely contribute to the preservation of our shared audiovisual memories from around the world.

The Digitization of Films and Photos of the Istituto Luce

Edoardo Ceccuti

Istituto Luce

Abstract

We truly think that if preservation is a clear imperative, making the heritage accessible is a moral imperative. This is why we made the decision of posting our heritage on Youtube—which does not mean abandoning our documents to the logics of web exploitation, often in contradiction with public policies in regards of documental preservation and assessment and certainly not comparable to the “depth” of experience guaranteed by an offer that includes the results of archival work. It will still be the job of our archivists, cataloguers, as well as the complexity of our database, to guarantee the mediation and ensure that a quick encounter can become a cultural and educational experience. Integration and the joint effort of archives, respecting copyrights and different identities, is a challenge for the creation and reinforcement of a mutual and shared memory. This resource is necessary for whomever wants to face the globalized world.

Author

Edoardo Ceccuti is the Director of Istituto Luce’s Historical Archive and Documentary Production Department. He has been in the film industry since 1973 when he joined Warner Bros to work as an executive on many European film productions. In 1980 he left Warner Bros Italy to join Gaumont Italia. Since 1988 he works for Istituto Luce.

Istituto Luce was the main means of information, education, and propaganda during the years of Italian fascism. Its original heritage—covering 1925-1945—included nearly 3,000 newsreels, 3,000 documentaries, 650,000 photographs, and 3,136 historical documental units. Through a complex production system, for twenty years Istituto Luce developed an integrated “media system”, that on one hand offered movies as an educational channel (such as the “Cinematiche” collection, an audiovisual series presided by a committee of experts that included instructions for farmers on how to increase the efficiency of their crops; or touring documentaries that described the beauty of Italian cities to Italians and foreigners); while on the other produced reports or propaganda newsreels as frequent as five a week during the years of consent. This entire organization was an absolute novelty in the mid-Twenties. Luce movies were obligatorily screened (due to a 1926 law) in all theaters of the Italian Reign, becoming the main educational vehicle in the country for the great Italian mass that at the time was mainly illiterate. Even all photography needed to be purchased obligatorily at Luce for publishing through main media.

With the fall of fascism, Istituto Luce maintained, in spite of a few institutional passages, its public nature, which allowed it over the decades to acquire most of the newsreels and documentaries produced in Italy in the decades between the end of the war and the Eighties. Presently, this makes it the main Italian archive for nonfictional films and documental photography. Today Istituto Luce preserves thousands of hours of film, and nearly three million photographs that altogether cover almost entirely the history of Italy of the Twentieth Century.

This great patrimony, preserved for decades thanks to traditional tools and supports, with the rise of digital, presented and presents us with new challenges. To begin with, because as a public institution delegated to the preservation of historical audiovisual documents, we are invested with a sense of

responsibility towards the future generations, for whom audiovisuals (as already obvious today) will be the main source of knowledge and communication. In second place, we have been dealing with the issues of preserving our documents through modern technologies since the very dawn of digital tools for moving images; and finally, because we believe that digital resources are fundamental not only for the preservation of culture, but also for the exploitation and valorization of culture—which explains why we not only focus on preservation, but also on the web and its resources.

We believe that the topic of this Session was perfectly targeted: “Practical challenges, theoretical solutions.” Our twenty-year old history of relation with digital has actually followed this path: we worked on finding theoretical solutions to practical problems and then, after finding them, we needed to translate them into practical solutions. In the audiovisual field (and not only), this means finding the most coherent answer, keeping in mind technological options (Moore’s well-known law is certainly an extraordinary opportunity, but also a problem for those who need to invest resources), as well as economical possibilities, while keeping in mind that the genre we work with (even when concerning digital) requires higher expenses in terms of technology and memory.

Speaking about the “challenge of digital preservation,” which is the main topic of this conference, there is fundamental matter that has always inspired our policy, and that can perhaps offer a common matter for reflection. Our archive was not “born digital”: it was founded in 1925 as a public film production company. In those years, internationally, film preservation was at its dawn, and films were still considered, rather than documents of memory, a means for entertainment. As known, today film is the only preservation medium for moving images, and it has proven to survive more than a century. Our film archive (besides a few accidents which caused the loss of some material), is almost entirely preserved in 35mm in air-conditioned cells. This activity clearly requires huge fatigue and resources, but let’s admit it: it allows us archivists to sleep tight.

So when challenged by the age of digital, we didn’t have the worries of who starts out with an ephemeral patrimony (we all know about archives born within the world of digital and lost forever). Actually, our problem was quite the opposite: how to share the richness we had, that was consulted only by a few dozens of film experts, and make it become public patrimony? The answer certainly lied in digital. So we started a vast project that included transferring the entire patrimony in low resolution, starting with films (and now also photography): nearly 100,000 documents—from a single newsreel report, to an hour-long documentary. After this, a complex database was built, and supplied with an advanced search engine. In the end, all documents were analytically catalogued.

This gave us the opportunity to understand that the real bond between digital preservation and exploitation/sharing is the database itself; a tool, or better, a true patrimony, besides being a sort of “guarantor” of the system. After all, Article number 1 of the “Charter on the preservation of digital heritage” is very clear on this point: “Digital materials include texts, databases, still and moving images, audio, graphics, software and web pages, among a wide and growing range of formats. They are frequently ephemeral, and require purposeful production, maintenance and management to be retained.” Actually, in spite of its digital nature clearly “ephemeral”, a database offers a fundamental surplus to traditional documents, whether born digital or eventually digitalized. Additionally, this can be considered a preservation method which not only doesn’t forge original contents, but rather emphasizes them and “conveys” them once again. Databases are new documental archives, and not for their content, but for what they make available; they create new meanings.

When it came to digital preservation, we put in act a policy that we felt was the most appropriate, keeping in mind the quality of the analogical documents that had been preserved till then on their original

media. Films and photography were acquired by scanner and saved into an international standard data format. The digital copy went through reconditioning and restoration and was then digitally archived again. The material could then be transferred automatically into new storage systems or transferred back into film, without any apparent loss of quality. The new generation of scanners, in our case two and four K, today allows us to transfer film and photography at a definitely lower cost compared to the past, although the quality of the transcription remains excellent. Over the past years, this patrimony was often reversed back into analogical, which allowed us to flank the traditional preservation with a second print on a safety support.

If preservation is a clear imperative, making the patrimony accessible is a moral imperative, as we've always asserted. All our newsreels, documentaries and a vast percentage of photographic material are accessible online thanks to an advanced search engine that gives access to catalographic data, while visualizing the images. For our cataloguing system we use *xDams*, created thanks to a project financed by the European Union. It is a specialized and integrated system for the treatment of diversified documents based on a common technology, description, and methodology. This platform, built on a total web-based procedure, uses open data format, such as XML, as a model of representation and format for data preservation; as well as OpenSource software such as JAVA 2, and Jboss Application Server, as the development system for multiplatform flow both on the client and the server's sides.

The variety of contents within our archive (that not only covers Italian political, economical, and social history, but also the history of other countries) requested the creation of a Thesaurus for research management, and today this Thesaurus includes nearly 70,000 entries, and can practically be considered an encyclopedia abstract of the Twentieth century. Access to the database takes place through a search engine based on entry, the consultation of indexes and dictionaries, structural navigation of each single archive source, and search engines shared through authority files.

As for accessibility, we are aiming to evolve the actual website in an editorial direction, as well as opening up to other digital contexts. At the moment, we are also working on offering research in different languages. Another of our main objectives is to offer licensing for educational and scholastic purposes, in line with the *Creative Commons* standards. Last but not least, we intend to offer the entire audiovisual patrimony in Europeana.

Over the years, we have often modified our guidelines, introducing new innovations—whether technical or cultural—brought by computer technology and by the web, all the way to introducing web 2.0. However, we've never distorted our institutional mission, which aims at offering our contents to the entire web community. Interacting with users, has, as a matter of fact, given us the possibility to broaden the knowledge of our own patrimony and put in act specific projects with different subjects. Once again, our database is our best ambassador: not simply a showcase, but a fundamental tool of a cultural policy that today allows us to be a core player, both on a domestic level (when dealing with other film archives) and on an international. It is in fact thanks to our database that we were included in several European projects.

Since we've launched our website in year 2000, its traffic numbers has always been gratifying, and the tumultuous evolution of the Web itself has forced us to reach out towards new objectives, while pursuing even more ambitious goals. If the foundation of our cultural policy is and remains the scientific cataloguing rigor and our search engine—as you've certainly understood—this should never become a hindrance or nonetheless an excuse not to experiment the new prospects offered by the Web.

Our database is founded on international archival standards and rules (ISBN), along with the audiovisual regulations (starting from Fiaf standards). Each document is described individually, and along

the data, there is also a detailed description of the content—by word (organized within the Thesaurus), and by sequence. This is a long and complex job—available to experts as well as to the common users—and it offers the possibility to obtain precise results and controlled information.

And this is why we made the decision of posting our material on Youtube—which does not mean abandoning our patrimonial documents to the logics of web exploitation, which are often in contradiction with public policies in regards of documental preservation and assessment. In order to do so correctly, we kept in mind a few facts. To begin with, most audiovisual material online is exploited through the video Google platform, which has developed a high-quality video system. Next to this, Google offer tools to third parties free from bonds, since its main advantage lies in the increase of volume and quality of its content. This choice is also supported by another general consideration: sharing and independent development of contents are the core nature of web 2.0, and Youtube is its most evident expression. Even those who pursue public policies, cannot set aside this reality, even more in the field of culture. For years now, youngsters have abandoned television to reverse their attention to web content, and it is mainly to these youngsters that we want to reach out, channeling their attention on the content of our historical patrimony.

The exploitation and search engine resources are necessarily different. As undeniably verified during the first weeks on Youtube, after posting 30,000 clips, we can assert that, thanks to these platforms, digitalized cultural content and even entire historical patrimonies receive public attention—and in particular the attention of the younger generations—that is not comparable to an institutional offer, quantity-wise. However, what about when it comes to quality?

This point needs further reflection, but I believe this conference is precisely the best place where to face this question. The exploitation of cultural content through a dedicated channel such as Youtube, is certainly not comparable to the “depth” of experience guaranteed by an offer that includes the results of archival work, as well as editorial. This matter refers to a long-time discussion, at least twenty years, considering the web’s age, concerning the role of traditional mediators in the age of digital. Digital today pervades schools, universities, publishing houses, and newspapers, and while doing so, these same institutions seem to lose their function due to technological innovations that put the user in immediate contact with documents and information, with the logics of web 2.0, or of “home-made” web products, that we have no particular reason to not consider worthy and “democratic.”

However, the fact that there has been an urgency for an international conference such as this to discuss the destiny of our digital memories, proves that both the market and the Internet—that follow their own policies and that are often incompatible with each other’s—cannot guarantee that our patrimony will not be misplaced, nor can they guarantee that, without a correct mediation, the users will obtain their cultural objective. So, as suggested by the premises of this conference, what truly matters is the correct balance between industry, university, and public policy, in order to make our digital resources the starting point for a renovated idea of patrimony and culture. Our agreement with Google should therefore be considered from this point of view: as the beginning of a dialogue between public institutions and private companies. And although both naturally pursue different objectives, both have decided to put together their own resources for a maximum distribution of culture and memory. Youtube represents a ground to meet with an audience and a generation unaccustomed to traditional archives and memory itself. Yet, it will still be the job of our archivists, cataloguers, and editors, as well as the complexity of our database, to guarantee the necessary mediation and ensure that a quick encounter as the one that takes place on Youtube can become a cultural and educational experience.

Before ending, I believe it is important to leaf through a few points that are among the main objectives of this conference. And I will share with you how we consider to proceed with a few of them.

- Regarding preservation proposals, I believe it would be important to open a discussion, as I just pointed out, concerning the relation between preservation and exploitation.
- As for matters concerning exploitation and copyrights, we need to remember that due to their commercial value, audiovisual documents (as photographic ones, naturally) are conditioned by the market, which at times can stiffen accessibility due to public policies, with negative consequences also on accessibility for studies/researches. Our choice for this point—and our agreement with Google is only the last chapter—is to consent free access to online consultation in order to never hinder research;
- As for exchanging chosen standards, we are clearly available to any initiative in this direction. And we will also put at disposal our experience in the fields of digital preservation, audiovisual cataloguing, and database, specifically;
- When it comes to specific competences, over the many years, our experience has led us to find and choose audiovisuals and university experts (archivists, historians, computer scientists, etc), each of which is essential to a definition of policies, that if not considered carefully, can lead to a serious loss of economical resources.

In the end, it is clear that over the years digital technologies and the Web have transformed our lifestyle, our daily life, and the way we plan our future—whether people like me, that still belong to the Twentieth Century, like it or not. However it would truly be ungenerous and nostalgic not to admit its advantages. Thanks to the web, today audiovisual digitalization not only offers accessibility to an endless number of users but most of all it integrates and creates a dialogue among documents coming from different archives, thus offering new value to the documents and more significant interpretations. In the same way, major archives have broadened research opportunities by sharing catalographic systems and constant database interaction, also thanks to standard thesaurus and intelligent predisposition of user maps.

Integration and the joint effort of archives, if respecting copyrights and different identities, is a great challenge for the creation and reinforcement of a mutual and shared memory. This resource is now absolutely necessary for whomever wants to face the globalized world.

Challenges and Triumphs

Preserving HD Video at the UBC School of Journalism

Adam Jansen

School of Library, Archival and Information Studies, The University of British Columbia, Canada

Abstract

This paper provides an overview of the InterPARES 3 case study on preserving HD Video at the UBC School of Journalism. A selection of the significant challenges, both technological and procedural, and the main recommendations are discussed, including: balancing the archival needs of preserving the video assets against the creative workflow and editing process; modifying the asset management software to incorporate elements of the InterPARES Chain-of-Preservation model; implementing the PBCore metadata schema; providing rapid, secure access to extremely large video files; and creating of a backup methodology to safeguard assets within the archives.

Author

Adam Jansen is an information management consultant pursuing his Ph.D. from the School of Library, Archival, and Information Studies at The University of British Columbia. He holds an Interdisciplinary M.Sc. from Eastern Washington University in Business Administration and Computer Science and was involved with the InterPARES research project, first as a co-investigator and then as a graduate research assistant, from 2008 to 2011. His research interests center around the creation, use, preservation, and forensic analysis of trustworthy digital records.

1. Overview

The University of British Columbia's (UBC) School of Journalism (SoJ) approached the InterPARES project team to request help preserving the documentary video footage created by the International Reporting Class. The SoJ has the distinction of being the only graduate level program in journalism in Western Canada, and is committed "to achieve the highest professional standards in journalism through instruction in journalistic practice and the scholarly understanding of journalism, critical thinking, and teaching of ethical responsibility."¹ In support of its mission, the SoJ has summarized its goals as:

- To produce a new generation of journalists with the specialized knowledge, cultural awareness and critical thinking skills needed to excel in journalism;
- To improve the practice of journalism through the education and training of journalists in scholarship, research and professional development; and
- To advance through rigorous research the understanding of the vital role of journalism in the public sphere and to contribute to the current body of journalism studies existing in Canada.²

¹ "Journalism – Degree Programs -- The Faculty of Graduate Studies -- Faculties, Colleges, and Schools -- Vancouver Academic Calendar 2012/2013," UBC Graduate School of Journalism, accessed 30 August 2012, <http://www.students.ubc.ca/calendar/index.cfm?tree=12,204,828,1183>.

² "About the Program: Mission Statement," Internet Archive, accessed 10 January 2010, http://web.archive.org/web/20070704014540/www.journalism.ubc.ca/mission_statement.htm.

This level of commitment has resulted in a series of collaborative projects, such as with world-renowned journalist Dan Rather on *The Dan Rather Project*, and an Emmy award for outstanding *Investigative Journalism* for the documentary on e-waste titled *Ghana: Digital Dumping Ground*.³

As a member of the Canadian Media Research Consortium (CMRC), a partnership between the journalism programs at the UBC Graduate School of Journalism, York/Ryerson Graduate Programme in Culture and Communications and the Centre d'études sur les Médias at Laval University, the SoJ recognized the advantages to providing other universities, journalists and the public access to the large number of digital source files created during the documentaries created by its faculty and students. Creating a centralized video archive of high-definition documentary footage would support the CMRC's mandate to undertake research in media and communications, with a focus on technological change; to promote collaborative research focused on Canadian issues; and to disseminate the research findings of its partners.⁴ The SoJ recognized that the authenticity of, and long-term access to, the video assets contained within the archives would be of great importance if the archives was to be considered a trusted source of documentary video material.⁵

The SoJ project fit well within the objective of InterPARES 3 to "translate the theory and methods of digital preservation drawn from research to date into concrete action plans for existing bodies of records that are to be kept over the long-term by archives, "and was added as a case study.⁶ The goal of the case study was "To establish a digital video archive of high definition video footage created by the SoJ's students; devise means to ensure the preservation of the raw footage of student projects; and create policies allowing for the footage to be used internally and externally."⁷ The case study focused primarily on the video assets produced by the International Reporting Class, while building into the design the flexibility to accept video assets from other classes and external donors in the future. The International Reporting Class is a yearlong course for second year graduate students enrolled in the Master's of Journalism program.

Students enrolled in the class study the history of foreign correspondence while focusing on modern-day documentaries that exemplify best practices. The design of the curriculum centers on planning and producing a documentary report on a chosen topic of international significance. At the end of the first semester, up to ten students, one professor and two adjunct lecturers move out into the field where they collaboratively create high-definition documentary footage on the selected topic from various locations around the world. Upon returning to UBC, the students import the footage from the source files⁸ contained on tape and/or hard drive into the video editing system. During the editing process, the source files are broken into sub-clips,⁹ portions of which are then intermixed and combined into the documentary itself using Apple's Final Cut Server (FCSvr) video editing system. Every documentary within FCSvr has

³ "UBC Graduate School of Journalism wins Emmy Award for Outstanding Investigative Journalism," UBC Graduate School of Journalism, accessed 4 April 2012, <http://www.publicaffairs.ubc.ca/2010/09/27/ubc-graduate-school-of-journalism-wins-emmy-award-for-outstanding-investigative-journalism/>.

⁴ "CMRC," UBC Graduate School of Journalism, accessed 1 Sept 2012, <http://www.journalism.ubc.ca/about/cmrc/>.

⁵ For the purposes of this case study, authenticity is ascertained through establishing a record's identity and demonstrating its integrity.

⁶ "Project Overview -- InterPARES 3," InterPARES, accessed 2 September 2012, http://www.interpares.org/ip3/ip3_overview.cfm.

⁷ InterPARES 3, "UBC School of Journalism: Final Report," InterPARES, accessed 9 March 2012, http://www.interpares.org/ip3/display_file.cfm?doc=ip3_school_of_journalism_final_report.pdf.

⁸ Source files are the un-edited raw media files as shot in the field, prior to the creation of sub-clips.

⁹ Sub-clips are discrete video clips of one particular theme or topic that are of one continuous shot.

a corresponding project file, which is an XML-based file with pointers to every sub-clip used in the documentary, the location of the sub-clips at time of rendering, the order of the sub-clips within the documentary, and which portions of the sub-clips were actually used. In the last step of the editing process, the project file is rendered (the specified portions of the selected sub-clips are combined in the appropriate order) into a final project—a feature length, single video file suitable for television broadcast. Course instructors use the final project as the basis for assigning marks to the students enrolled in the class, and distribute the file to various news agencies for possible airing on their network.

2. Challenges

The first major challenge addressed in the case study was to define which video assets—source files, sub-clips, interim clips, project files, final projects, etc.—were to be stored within the HD Video Archives. The Dean of the School, the Chair of the Graduate Program, and the GRAs from the Archives and Journalism programs all possessed different views of: what the ‘asset’ was, the purpose of the HD Video Archives over the long-term, and the types of material that needed to be stored within the archives. Over the three years of the case study, the team continued to revisit the issue of what material ‘should be’ in the archives, with slight evolutions in the definition with each successive discussion. Determining which assets should be placed within the archives was complicated by the fact that, due to the collaborative nature of the final project drawing from a large number of sub-clips shot at various locations by any number of the participants in the class—student and instructors alike, determining the authorship of each sub-clip was virtually impossible.

Other challenges were presented in the form of specific constraints in the design of the HD Video Archives based on the existing practices and capabilities of the SoJ.

2.1 Artistic constraints

The SoJ considered the media assets stored within the archives as recyclable assets; that is, the intention from the beginning was that the assets in the archives were to be re-used for future works. The SoJ was less concerned about where the sub-clips came from and how they were used, but rather focused on how those clips could be used on other projects; as their view of the primary purpose of the archives was to centralize, organize, and preserve the assets for future journalistic pieces, by UBC students or external journalists. As such, the ‘recordness’¹⁰ of the sub-clips was of a secondary nature to the SoJ. The only pieces contained within the archives considered inviolable were the finished clips produced by the students and other completed works donated from outside sources.

The workflow constraints on the use of the media assets were also clearly articulated: any policies or procedures developed should not interfere with current workflow processes, add undue burden to the students’ workload, or hamper the inclusion of the media assets in future productions. The archives is viewed by SoJ as a means to quickly and efficiently locate documentary footage that has been preserved in a way that ensures its authenticity in order to produce artistic pieces of journalistic expressions along a multitude of topics. The software used within the SoJ to create the documentary works, in this case

¹⁰ Anne J. Gilliland and Philip B. Eppard, “Preserving the Authenticity of Contingent Digital Objects,” *D-Lib Magazine* 6, no. 7/8 (July/August 2000), accessed 25 August 2012, <http://www.dlib.org/dlib/july00/eppard/07eppard.html>.

Apple's Final Cut—a product that in 2008 accounted for a 49% market share—dictated the workflow of how the assets are used.¹¹

2.2 Technology constraints

Due to the limited budget at the SoJ, any recommendations had to work within the limited grant funds available while leveraging the existing technology environment to the extent practicable. Along with the budget constraints, the SoJ also had very limited technology support available to it. Most tech support for the School is received through the UBC Department of Technology, which provides “IT-related strategy, applications, infrastructure, and support services to the UBC community.”¹² While experienced at providing desktop and network technical support, UBC IT has little experience in supporting the Final Cut suite of products. Combined with the proprietary nature of the Apple product, there is a cap on the amount of customization that is possible to the existing software, both from an infrastructure/programmatic perspective, as well as procedural.

2.3 Legal constraints

The SoJ is an academic unit with the UBC Faculty of Arts and therefore is governed by the policies of the University and the admissions and curriculum requirements established by the UBC Faculty of Graduate Studies. As such, the work produced by the students stored within the archives are more than simple artistic works of journalistic expression—they are also records in the classic sense¹³ in that they are evidence of the SoJ's execution on its mission¹⁴ as well as evidence supporting the students' marks received in the course. The fact that the final projects are ‘records’ places the additional requirement on those particular pieces of student work that they be managed in accordance with the approved UBC retention schedule. The SoJ is also subject to the requirements of the *Freedom of Information and Protection of Privacy Act*, R.S.B.C. 1996, c. 165 requiring disclosure of those records maintained by the SoJ not explicitly exempted. UBC Policy 88 states that students retain intellectual property rights over the works that they produce, which mirrors the intellectual rights and moral rights laws. In Canada, while the intellectual rights can be assigned to third parties, the moral rights stay with the creator.¹⁵ As mentioned earlier, due to the collaborative nature of the works created at SoJ, ownership over the materials—and by extension the SoJ's right to place the material online for public access—required review by UBC legal counsel specializing in intellectual property.

2.4 Resource Constraints

The need for a HD Video Archives was not recognized in time to submit requests for additional funding, staff or resources in the current budget cycle. As a result, the SoJ had extremely limited funds and staffing

¹¹“Final Cut Pro Apple of Oscar's Eye,” C/Net News, accessed 3 March 2012, http://news.cnet.com/8301-13579_3-10465202-37.html.

¹² “About UBC Information Technology,” UBC IT, accessed 15 August 2012, <http://it.ubc.ca/about>.

¹³ “A document made or received in the course of a practical activity as an instrument or a by-product of such activity, and set aside for action or reference.” From the InterPARES Terminology Database, accessed 4 September 2012, http://www.interpares.org/ip2/ip2_terminology_db.cfm.

¹⁴ UBC School of Journalism Academic Calendar, op. cit.

¹⁵ Luciana Duranti, “The long-term preservation of the digital heritage: a case study of universities institutional repositories,” *Italian Journal of Library and Information Science* 1, no. 1 (2010): 160.

that it could dedicate to the development of the archives. What funding they did have available was received through a grant to support the preservation of the International Reporting Class's documentary footage—the grant was from the same source who donated the funds for the creation of International Reporting Class. The donated funds allowed the school to hire a technical consultant to install the FCSvr system; and, fortuitously, this consultant was able to arrange the donation of additional servers and a Storage Area Network (SAN) from a third party. The collaboration with InterPARES supplemented the limited staff and provided the much needed knowledge and experience in digital preservation. Toward the end of the project, the SoJ hired two additional summer interns to assist with the importing the 2009 and 2010 International Reporting Class footage into the archives. After the first summer, one of the interns continued on the project for eight months until the funds ran out.

3. Creation of the HD Video Archives

After a series of meetings to establish the project goal, expected outcomes, operating constraints, and timelines, the GRAs conducted an extensive investigation into the workflow used by the SoJ, the technical infrastructure available, and the current policies and procedures used by the students. The GRAs then researched journal publications, the findings from InterPARES 2,¹⁶ and the practices of several other video archives and news bureaus in order to determine current best practices for preserving digital video in a production environment. From this baseline, the GRAs created a series of recommendations for the creation and management of a HD video archives.¹⁷

3.1 Define assets to be stored

Through three years of the case study, the types of materials that were to be stored in the archives evolved as technology, policies, and staffing changes presented new opportunities to open the archives up to a wider definition of what material was appropriate for inclusion. To provide a consistent approach to the selection of materials for inclusion in the archives, the SoJ formally defined what the HD Video Archives was, including its mission, targeted user group, and the types of materials that it would be accepting for inclusion into the repository. Having this definition in written form provided a consistent application of the rules for inclusion across the School, the many classes, and the students who perform the work. This definition is static; rather it should be periodically reviewed for expansion or clarification, as needed by the introduction of new hardware, software, or new partnerships/donors.

3.2 Establish required authenticity and preservation metadata

Maintaining the authenticity of the media assets in the archives over the long-term requires that a very specific set of metadata be captured and/or created at the time the media files are produced and expanded when the assets are added to the archives. InterPARES2 studied the metadata necessary for long-term preservation as detailed in the Chain of Preservation (COP) model. The COP model describes the

¹⁶ InterPARES 2, *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experimental, Interactive and Dynamic Records*, ed. Luciana Duranti and Randy Preston (Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008).

¹⁷ For a more comprehensive description of the findings and recommendations, see the “UBC School of Journalism: Final Report,” op. cit.

lifecycle of a record from creation through final disposition, i.e., destruction or permanent retention in an archives. From this model, the team created an extensive list of metadata pertaining to the authenticity of electronic records over the long-term.¹⁸ In addition to archival and authenticity metadata detailed in the COP model, the GRAs heavily researched the PBCore and MPEG-7 metadata models used extensively in the media industry.

It was determined that the MPEG-7 model was too complex for the SoJ's purposes; to implement such a metadata model would violate the constraint of an efficient workflow production system. PBCore, on the other hand, matched a number of the COP elements and is a widely adopted metadata standard developed for the public broadcasting sector. The PBCore metadata schema allows for a high level of interoperability amongst public broadcasting stations and therefore strongly mirrored SoJ's desire to share the assets contained within the archives with other journalists and the public. The hierarchical arrangement of content classes, containers, sub-containers and, finally the elements within PBCore further allowed for the level of customization necessary to accommodate both the industry metadata elements and the archival metadata elements required. The content classes are "created as 'conceptual wrappers' that cluster together a list or structure of thematically-related Elements (metadata fields and their attributes and properties)."¹⁹

The four content classes allow for grouping the metadata elements into intellectual content (unique information about the asset such as title, unique ID, creator, etc.), intellectual property (owner, copyright holder, usage rights and restrictions, etc.), instantiation information (date created, aspect ratio, frames per second, video format and resolution, etc.) and finally the PBCore extension (additional metadata requirements that have been crafted by organizations outside of the PBCore Project.)²⁰ This last content class allowed for the inclusion of the COP metadata elements into the overall PBCore structure—allowing for the capture of the necessary authenticity and preservation metadata while still maintaining a high level of interoperability with other journalism organizations. The full list of metadata elements used in the HD Video Archives is in Appendix C of the Final Report.²¹

3.4 Establish Ownership over the assets in the archive

Establishing the intellectual rights ownership over the assets within the archives required a two-pronged approach: one for legacy materials and one for future materials. Working with UBC's legal department experts on intellectual property, the SoJ determined the appropriate measures to take in order to establish ownership of the video assets created in years past, while concurrently creating a licensing agreement for future SoJ students that will assign the ownership of material created to UBC. Even with the license in place, the issue how the use of those video assets by third party journalists will or will not conflict with the moral rights of the students creating the sub-clips has yet to be resolved. Additionally, several donors have also expressed interest in donating documentary video material to the SoJ for inclusion in the

¹⁸ InterPARES 2, "Appendix Fourteen: Chain of Preservation Model Diagrams and Definitions," in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES)2: Experimental, Interactive and Dynamic Records*, ed. Luciana Duranti and Randy Preston (Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008), accessed 6 August 2012, http://www.interpares.org/ip2/display_file.cfm?doc=ip2_book_appendix_14.pdf.

¹⁹ "Background of the PBCore Public Broadcasting Metadata Dictionary Project," PBCore, Corporation for Public Broadcasting, accessed 4 March 2012, http://www.pbcore.org/PBCore/PBCore_background.html.

²⁰ Ibid.

²¹ "UBC School of Journalism Final Report," op. cit.

archives, and by extension reuse by SoJ students, other journalists, and the public at large. Given the potential diversity of sources and unknown provenance of some of the material within the archives, the ownership—or lack thereof—is described in the metadata schema along with the usage rights according to the designations created for the Creative Commons licensing matrix.²²

3.5 Document current practices

Early research into the workflow process used at the SoJ between 2008 and 2009 found that although a workflow process theoretically existed for the International Reporting Class, it was determined that the students enrolled in the class were not following the expected workflows. Each student used the methods and processes that were most comfortable for them, resulting in a wide variety of storage locations, disparate folder names and folder structure, and naming conventions unique to each student. It was determined that this inconsistency was due, in major part, to the lack of a consistent training method and no identified trainer responsible to communicate the expectations to the students. Given that each subsequent year a new group of students is selected for the International Reporting Class, a large percentage of the group of individuals creating, describing and adding content into the archives possess no experience in the process. Such diversity in location and naming of the media assets would quickly make locating material in the archives extremely difficult. To mitigate this individualism in management philosophy, each student is now trained according to a workflow that details the expectations, policies, and procedures that they will need to be familiar with before starting their respective documentary projects. For the sake of consistency, the entire process was documented—with both textual descriptions as well as screen captures for illustration—for each type of camera that was used for the class. By the end of the case study, three generations of cameras were used, requiring three separate sets of documentations, as each camera captured and outputted the video files differently.

3.6 Assign staff to manage and oversee the archives

The students enrolled in the International Reporting Class create a majority of the content of the archives; meaning that every year a new crop of students will be creating, indexing, and managing more assets that need to be included in the archives. This presents two challenges to maintaining a trustworthy archives over the long-term: first, students, in general, tend to be more focused on satisfying the requirements for their class than they are on accurately describing assets for future use by parties unknown; and second, every year begins a new group of students with little to no experience describing digital assets, resulting in often very divergent approaches to indexing amongst themselves and from the procedures manual they are *supposed* to follow. While the instruction and documentation discussed above can overcome these challenges to some extent, they are not a panacea to the problem.

To provide a check on the work of the students, as well as to refine the current policies and procedures, the SoJ hired project staff to perform quality assurance checks on all the new assets added to the archives. The QA staff, having the most familiarity with the system and indexing criteria, has the reasonability to train the students. Prior to the class going into the field, the QA staff reminds the students of the accepted methods and provides them with copies of the documented procedures to take with them. Once the students return, the QA staff conduct checks on the assets that were added into the system—

²² Tama Leaver, “Creative Commons: An Overview for Educators,” *Screen Education* 50 (Jan 2008): 38-43.

making corrections as necessary and additional training as required. The QA staff members—previous students who have already experienced the process start to finish and free-lance journalists who can come back year after year—also have the primary responsibility for processing of the backlog of assets that have accumulated over the years.

3.7 Provide necessary resources for rapid access, expansion

Given that the primary purpose of the HD Video Archives is to provide long-term trustworthy access to documentary material for reuse on future projects, and given that these assets can be several gigabytes in size, it is important to the success of the project that the system be able to move these assets from the storage subsystem out to user—whether they are on campus or across the globe—quickly and accurately. Providing a robust user experience requires implementing sophisticated technology, which is at odds with the low budget, low-staff constraint established by the SoJ. Moving large video files around from storage devices to editing machines is best-accomplished utilizing Fibre Channel protocols through SAN subsystems. Fortunately, a large video production house in Vancouver donated some older SAN equipment to the SoJ.

In addition to providing fast access to large files, the SAN system will also allow for a relatively seamless expansion of the storage pool following a growth-on-demand approach. That is, by starting with a relatively modest four terabytes of SAN storage, the SoJ will be able to add additional terabytes of storage by purchasing storage enclosures without having to purchase additional servers or storage controllers—equating to a lower *per terabyte* cost moving forward. Due to the speed of the Fibre Channel protocol, standard Cat-5 wiring does not have sufficient bandwidth to maintain pace with the storage sub-system. To ensure the fastest transfer possible of the video files from the archives to the video editing stations located through the building, SoJ installed optic cables at key points. While this provides a robust user experience within the building, external users are still limited by the capacity of their internet provider.

3.8 Develop backup policies and procedures

The urgency for a robust backup of the HD Video Archives is two basic issues: first, equipment fails—donated electronic equipment that has already been heavily used more so—and best practices recommend creating a backup to protect against any single point of failure; second, FCSvr is a proprietary system—meaning that subsequent versions may not be backward compatible or the vendor may, at some point, stop supporting it altogether. As part of the backup policy, the assets in the archive are mirrored onto a second set of hard drives to protect against the failure of any single hard drive or storage component. A tape drive was acquired with appropriate back up software, but funding has limited the number of tapes that could be procured. In the short term, this constraint is limiting the extent to which the entire system can be safe guarded against localized events (fire, flood, earthquake, etc.). The preferred backup strategy would entail both local backup through replication to a secondary storage array, as well as a tape backup that is maintained at a location 30kms or more away from the primary system to protect against data loss due to localized disasters. In the next budget cycle, the SoJ will be requesting additional funds to extend the backup policy to allow for both quarterly full backups as well as weekly incremental backups. Backup policies are only as good as their execution; it is important that the backups be routinely screened for accuracy by restoring the backups and comparing the results to the originals. This strategy has the twofold

benefit of ensuring that the backup policy accurately protects the specified material²³, and providing the IT staff with the familiarity of how to quickly and smoothly conduct the backup and restore procedures in the event that the system needs to be restored.

3.9 Export data

Maintaining the assets exclusively in a proprietary system places said assets at risk of becoming inaccessible as time and technology render the software platform obsolete. Based on the findings of the previous InterPARES projects, the SoJ was encouraged to maintain a copy of the original media assets, along with the corresponding metadata uniquely tied to the original asset, outside of the FCSvr system. The reasoning behind this recommendation was that in the event that the software platform becomes obsolete, is no longer supported, or the SoJ moves to a different software/hardware/technology platform, the original assets could be migrated onto the newer platform along with all the metadata that was created for the asset.

3.10 Use standardized naming conventions and descriptions

Unique file names are necessary to differentiate one media asset from another. While this function can be handled by the database contained within FCSvr through the use of GUIDs,²⁴ the GRAs on the project recognized two scenarios where the media assets themselves needed to be human recognizable. The first related to the fact that one of the primary purposes of the archives is to provide public access to the assets in the system. As it is unlikely that a majority of the external users will have FCSvrs of their own, these users will have to export the assets from FCSvr and store them locally. Outside the FCSvr system, the GUID file nomenclature would be of little use to the journalists. The second scenario centered on the goal of this case study to provide long-term access to the media assets created and stored within the archive. The GRAs determined that reliance upon a piece of proprietary software to maintain the link between file and metadata was not in the best interests of the project. One way to avoid this reliance on a proprietary software intermediary is to create a file name that makes sense to the people using the files.

Guided by Anne Thompson's work on standard naming conventions,²⁵ the GRAs recommended the naming of individual media assets using the following elements:

1. Course number
2. Project name
3. Date
4. Sequence number

Following the recommended nomenclature allows for a basic understanding of the context of creation of the asset without having to open the file itself. Additionally, two other key indexing fields within the FCSvr workflow provide further information about the media assets in a standard format.

²³ That is, the backup copy is able to completely reproduce those files it was instructed to backup and that the restored files are exactly the same as the originals that they are reproducing.

²⁴ Globally Unique IDs – GUIDs are 128 bit values typically stored as a 32 bit hexadecimal values allowing for 2¹²² possible values that can be assigned.

²⁵ Anne Thompson, "Standard Naming Conventions for Electronic Records," accessed 14 November 2009, http://www.sfu.ca/archives2/rm/rm_fundamentals/07UKFileNameConventions.pdf. The naming conventions recommended were developed according to the format outlined in this document.

For the description field, the description starts with the type of clip it is—interview, b-roll, A camera, B camera, etc.—followed by the formal name(s) of the interviewee(s), the name(s) of the interviewer(s) and the primary subject or purpose of the shot. The location is described first with the formal name of the location of the shooting (e.g., name of hospital, park, or place of business), the city where it is located, the state or province, the country, and the date of creation in ISO format.²⁶ The file naming convention combined with the description format provides sufficient information about the file to provide users a basic understand of what the context of the file, such as when and where it was shot, the names of the main parties involved in the footage, and the type of footage it is. Most importantly, the video assets are identifiable outside of the FCSvr system; providing a limited failsafe in the event that the need arises to replace the Final Cut product.

3.11 Migrate to supported products

Within information technology, the only surety is change. Moore's law states that speed and capacity double every eighteen months,²⁷ and as a result, the lifecycle of software products averages two to four years. With the introduction of new products, older products are no longer supported. On average, new versions of major software products are released every two to three years, with technical support typically available only for up to the three previous versions.²⁸ This approximately equates to a ten year window as a theoretical maximum in which to use software before it is no longer supported. This window can be much shorter based on major changes to hardware and operating systems, such as the migration from 32-bit computing to 64-bit computing. It is in the best interests of the archives that the SoJ maintain the system using vendor supported versions of software—resulting in an software upgrade every two to six years in order to stay within a support window.

Without vendor support, the archives will become increasingly more difficult to maintain and potentially become incompatible with future technology releases. A lack of vendor support will require the SoJ to keep the current generation of software and hardware operational indefinitely. Over the short term this is a plausible approach to the problem, but as the file sizes continue to increase exponentially (such as with the adoption of the 4K family of digital video)²⁹ and hardware migrates fully to the 64 bit computing platform, the current generation of technology may not be able to handle the increased demand put on the system. To maintain the compatibility of the existing system with newer technology, the SoJ would need to conduct extensive testing of the newer technology prior to introducing any changes to either the hardware or the software platforms. At some point in the future, it is highly likely that FCSvr will simply no longer function on the latest generation of technology.³⁰

²⁶ International Standards Organization, ISO 8601:2004 Date Elements and Interchange formats – Information Interchange – Representation of dates and times. 2004.

²⁷ G.E. Moore, "Cramming More Components Onto Integrated Circuits," reprinted in *Proceedings of the IEEE* 86, no. 1 (Jan 1998): 82-85.

²⁸ "Office Family Product Support Lifecycle FAQ," Microsoft Corporation, accessed 14 August 2011, <http://support.microsoft.com/gp/lifeoffice>.

²⁹ Current HD technology utilizes 1080 lines of resolution, each frame equivalent to a two-megapixel photo. The large-budget movie industry is switching over to the newer 4K platform that captures 4096 x 3072, or each frame having the equivalent of a 12.6 megapixel photo.

³⁰ For example, WordStar and Dbase, two widely adopted programs (for word processing and database management, respectively) from the 1980s ceased to function on the x86 platforms of computers of the 1990s.

4. Conclusion

The SoJ requested InterPARES3 to “research, create, and implement a plan to preserve and index a high definition digital video archive in online and electronic formats.”³¹ What resulted from this case study was an in-depth analysis of the existing workflows, policies, procedures, training, technology, and staffing used in the International Reporting Class. Based on the analysis and extensive research into existing best practices used by major news bureaus, a series of recommendations were presented to the Dean of the SoJ and a majority of the recommendations was put into practice. Along with the recommendations, the GRAs created procedures manuals for each type of camera being utilized in the class, the UBC legal department developed a licensing agreement for each student to sign before going into the field, and a new metadata schema was developed that would allow for the capture of management, archival and preservation metadata.

By the end of the case study, the SoJ under the direction of the InterPARES GRAs installed an entirely new hardware and software system. These upgrades expanded the capabilities of the SoJ to provide UBC students and the public access to a growing collection of video assets while providing increased system response, a noticeable improvement in the movement of the files across the network, greater storage capacity, and remote backup of the assets and their associated metadata. Rules of operation were created within the FCSvr workflow system allowing for rights based access to active and donated media files, identity management allowing for the identification of file creators and editors, and controllers over the editing and deletion of inactive media assets (such as student’s final projects and donated material). Lastly, the design of the network allows for the seamless addition of servers or storage following a growth-on-demand philosophy; allowing the SoJ to maximize its limited funds through targeted hardware procurement.

Shortly after the case study ended, the first great challenge to the longevity of the archives arose. One of the ‘worst case scenarios’ came to light—Apple announced that the Final Cut Server software upon which the archives was built was going straight to end-of-life.³² Dealing with this scenario will prove a robust test to the veracity of the recommendations implemented by the SoJ. When a product reaches end-of-life, Apple will no longer provide any support for the product, nor would they test the compatibility of the legacy product with upcoming Apple Operating System software or provide patches/fixes/updates to any issues that arise. The current strategy at the SoJ for addressing this issue is to keep the current generation of hardware and software functional for the foreseeable future, while addition funds are requested to upgrade the entire system to a newer platform. If funding can be obtained, a new round of research will be conducted into the capabilities of current the video editing systems to meet the needs of the SoJ, its students, its mission, and its ability to continue to share its invaluable media assets with journalists and the public around the world. Until such a time as the migration to a new platform becomes feasible, the SoJ will continue to back-up the images and export all of the metadata via XML files in preparation for migrating the same into the new system.

³¹ “UBC School of Journalism Final Report,” *op. cit.*

³² “Final Cut Server has been discontinued,” Apple Corporation, accessed 14 Mar 2012, <https://discussions.apple.com/thread/3131590?start=0&tstart=0>.

Acknowledgements

The SOJ case study was far from the work of any single person. Over the three years of the case study, ten GRAs (including the author) contributed to the research, findings, and recommendations that went into the final report. This proceedings paper would not have been possible without their efforts, energy, and enthusiasm to see this project through and the author would like ensure they are recognized: Karine Burger, Kerry-ellen Canning, Donald Force, Judy Hu, Donald Johnson, Adrienne Lai, Shamin Malmas, Karen Moxley, and Elizabeth Walker. Guiding the efforts of the GRAs were the project coordinators, Randy Preston and Alexandra Allen, while kind support was offered by Sandy Orr and the results disseminated by our webmaster, Jean-Pascal Morghese. Lastly, none of this would have been possible without the tireless work and dedication of the InterPARES director, Dr. Luciana Duranti, whose vision, passion, and drive was an inspiration and source of awe to ten years of GRAs through the three phases of InterPARES.

Digital Disaster Recovery for Audiovisual Collections

Testing the Theory

Mick Newnham, Trevor Carter, Greg Moss and Rod Butler

National Film & Sound Archive of Australia.

Abstract

If an audiovisual collection is to be preserved and fully accessible then digitization is really the only option. However, digitization is expensive and has a short format life compared with most analogue audiovisual media. Consequently hard financial decisions about the way a collection is to be managed into the future need to be made by those responsible. Risk management is a crucial part of the decision making process. This paper outlines and follows the logical steps that may be taken to improve the ability of an organisation in preventing and recovering a digital collection from a range of disasters.

Authors

Mick has been involved in audiovisual preservation since 1988 and is currently the Manager of Conservation, Preparation and Research at the National Film and Sound Archive of Australia. Mick has been active in contributing to the work of the South East Asia Pacific Audio Visual Archives Association (SEAPAVAA), the International Federation of Film Archives (FIAF), Association of Moving Image Archivists (AMIA), International Centre for the Study of the Preservation and Restoration of Cultural property (ICCROM) and the International Organization for Standardisation (ISO). Mick has worked with collections in SE Asia, the Pacific, the Caribbean, Latin America, North America, Africa and Europe providing training and general consultancy services regarding preservation and collection management. Mick is a course developer and tutor with the Charles Sturt University on-line course in audiovisual preservation.

Trevor Carter is the Digital Migration Officer for the Motion Picture Laboratory and the National Film and Sound Archive of Australia with primary responsibility for managing the data sets associated with the collection of digital feature films and researching equipment and procedures to handle this material.

Greg Moss is the Manager of Digital Systems at the National Film and Sound Archive of Australia (NFSA) based in Canberra, Australia. This role has primary responsibility for the implementation of the NFSA's digital platforms and workflows within the Mediaflex MAM system, and control of the NFSA's digital Archive system. For almost 20 years Greg was the Audio Electronics Engineer at the NFSA responsible for the design and construction of the NFSA's audio facilities and digitization programs. He is also a member of the Society of Motion Picture and Television Engineers (SMPTE), the Audio Engineering Society (AES) and a member of the International Association of Sound and Audiovisual Archives (IASA) technical committee.

Rod Butler has worked at the NFSA for more than 20 years in a range of collection access, digital management, conservation, and preservation roles. Currently he is the head of the Preservation and Technical Services Branch which preserves the NFSA's audiovisual collection through duplication and conservation services. Rod has a strong focus on maximising the NFSA's ability to deliver prioritised, balanced and high quality preservation and access outcomes, and has delivered presentations and workshops on this topic at a number of international conferences.

1. Background

Audiovisual media have recorded the 20th century in way no other era has been recorded previously. Film, audio and video have enabled significant people and events to be witnessed by millions of people and it is hard to imagine a world where moving image and recorded sound did not exist. The problem of preserving this amount of information in the original analogue formats has been monumental and despite the best efforts only a fraction of the original recordings made survive worldwide. The skills required to adequately preserve and make accessible the remaining records have been honed for only the past two decades. And now the world has moved into the digital realm.

The quantity of data created by even a small digital audiovisual collection is massive by any measure. Individual files may be in excess of 10 terabytes!¹ This has created a new set of problems and demanded audiovisual archivists acquire a new set of skills while still requiring the original skills to manage the legacy collections. The costs required to digitize a legacy collection are largely beyond reach of all but the best resourced archive, and yet this is required if a collection is to be preserved and accessible. Consequently hard financial decisions about the way a collection is to be managed into the future need to be made by those responsible. Risk management is a crucial part of the decision making process.

Digital preservation risk management practices make use of a variety of strategies including the obvious legacy solution carried across from analogue collection management such as multiple copies and geographically separated storage. For many small collecting organisations risk management means a second data tape stored in a different part of the building as this is all that may have been budgeted for; and it appears to satisfy the requirements of managing the risk.

However good risk management of a digital collection requires a great deal more than a second copy. A disaster should be thought of as any incident or event that may potentially prevent permanent access to the record. As such an effective disaster plan must encompass a full regime of assessment, checks and testing. This requires the full support of the entire organisation from the top levels down and can be best managed when disaster mitigation is thought of as equally important to a collecting organisation as access.

Digital collection disaster management is a continuum starting with ensuring the original file is intact to begin with by a thorough quality checking procedure, the plan follows through with strategies for managing the changing environment of files types and hardware evolution and minimising the potential for loss by negligence or malicious attack and onto reducing the impact of catastrophic disasters on a regional scale.

2. Risks

To determine the best way to approach a disaster recovery is to act proactively and mitigate the impact on the collection as much as is possible. The first step is a thorough analysis of the risks that may affect the collection, the probability of each and the impact of each type of disaster. The following information has been collated from published sources on the internet^{2, 3} and most accurately refers to the experience of the

¹ Motion picture films digitized at 4K.

² J. Lainbart, S. Robinson, and M. van Zanelhoff, "Managing Threats in the Digital Age: Addressing security, risk and compliance in the C-suite," IBM Global Business Services, Executive Report, 2011.

³ "Causes and Cost of Data Loss," Any Software Tools, <http://www.anysoftwaretools.com> (accessed 21 February 2012).

Example 1

A 10TB system could store approximately 420 hours of PAL standard definition video losslessly compressed in Motion JPEG2000

Assumptions:

Bit depth	8 bit
Average data rate	55Mb/sec.

USA and European researchers for a typical business IT situation. These may be considered as *generic* factors. Audiovisual archives, where much of the digital collection will be as a result of the digitization of analogue materials to less ubiquitous formats, can add several other risks to be managed.

The original digitization of the audio, video or film requires close scrutiny to ensure that there has been no corruption of the content introduced when compared to the original source. It is important to distinguish between blemishes/artefacts that may have existed in the original and any that appear on the digital duplicate. Apart from the obvious issues of compression artefacts due to poor or compromised file format choice, degrading noise or dropouts may be introduced by some minor issue within the network or transcoding errors from the raw data to the final file format. Therefore a crucial step in considering a recovery plan is to ensure that any recovery is not attempting to repair existing issues.

The large quantity of information contained in an audio visual collection requires any digital archive to be a comparatively large digital storage system. Systems in the order of 5-10 terabytes would struggle to hold even a moderately sized collection of standard definition video, even using losslessly compressed video files.

Given the rate of change in storage technology a close watch on how new generations and shifts in technology is required to ensure that support can be maintained. The frequency that would be required to maintain technological currency is not fixed but will depend on the global conditions and market. Failure to maintain the collection with formats and equipment that still has technological support raises the serious and highly probable risk that even a slight system malfunction has the potential to cause significant or even total loss of the collection.

The comparatively high costs of migrating a digital collection to a new format/hardware system means that for many collecting organisations there will be the temptation to push the frequency of the migrations to the limit of support. As the end-of-life of the format/hardware approaches there is an increased risk of being caught out with a sudden change in support if the global market shifts faster than predicted.

Reliability of the power supply is a major factor. If there are interruptions to the supply during crucial file writing stages data will be lost or corrupted. The level of file corruption may be slight, but if it is not detected before the checksums are added further routine checking will not report an error.

2.1 Mitigation of risks

The basic principles of organisational disaster planning apply to all collecting institutions regardless of whether the collection is analogue/object based or digital. The overall risks such as fire, flood and security all need to be addressed at the organisational level.

The potential disasters require more than identification. Most disaster guides use a matrix of potential and impact to rank the risk presented by the disaster to create a priority listing of disaster mitigation and recovery strategies. A highly likely factor, for example water entering the building from

poor guttering, may only have a very minor impact on the collection, whereas a factor with a low probability, for example a major fire, would have a major impact.

Additional digital collection risks require additional mitigation strategies; however this should be considered part of the overall preservation strategy and incorporated in the workflow.

In the figure above, following the workflow from the initial digitization step; after the digitized content passes the quality check step a *checksum*, also known as a *hash sum*, is added to the file. A checksum is a numerical value calculated in a specific way from the bits in the file. At any point when the accuracy of the file needs to be checked, for example in a migration from one server to another, the checksum value is recalculated and compared with the original. If the figure match then the probability of the file be uncorrupted is very high. The actual value is unique to the file and will change if the file is edited or migrated to a new format.

There are many different algorithms or methods of creating a checksum. The simplest form is a longitudinal parity check where the bits are broken into “words” with a fixed number of bits. The words then are used to calculate an *exclusive or* function for the file.

Case Study: Using checksums to locate corrupted files

A series of born digital video files from a popular television show totalling 3.2TB arrived on removable media. The connection speed of USB2 meant that the transfer took several days to complete. The process appeared to have proceeded smoothly as no errors were reported by the Cycle Redundancy Check (CRC), a standard method used when copying material on hard drives and networks to insure error free copying. A comparison of the properties as reported by Windows XP showed the exact number of files and reported the same disk space was taken up by both the original and new copy of each file. As material had up until this time arrived in such small numbers no automated quality checking (QC) functions were in use. Having someone watch of each episode would have taken around 100 hours to complete and, as there was no suggestion of problems, was felt to be unnecessary.

To make sure the files codec used in the QuickTime wrapper were supported by the system 5 episodes were picked at random and played back in real time. While watching the 4th of these files an odd artefact was noted toward the end of the file.

< slice error.tif>

This artefact was just on the one frame and was thought to be a transcode error made at the time of producing the backups. To let the producer’s know about this issue the original file was played to compare with the copy. In this process it was identified that the copy had the artefact but the original was fine. The original file was copied again and the new copy was fine. This appeared then to be a random error, even so the likelihood of this occurring again was high. At this point it was decided to use a checksum to confirm that the information was correct.

Once the checksum was applied 7 files were identified as having problems. When these seven files were copied across again 5 were still found to have errors, however the errors were occurring in different location!

In order to achieve no errors the process of copying and validating had to be done a further three times to insure all files had been copied without errors.

Testing showed that it was a faulty USB cable that was causing the problem but it clearly illustrated the point that a checksum was the only way to insure data validity.

An important component of any preservation strategy is to spread the risk across several copies of the content. In the analogue environment this may have meant creating several duplicates of the original tape, disk or film. Each analogue generation would be slightly degraded in technical quality due to inherent losses in the transfer, however this was unavoidable and was accepted. However in the digital domain additional copies may be regarded as identical. The crucial consideration for the additional copies is that each is checked for errors prior to committing the file to the Media Asset Management (MAM) system.

The number of digital copies is not standardised however as a rule of thumb most major organisations create three full resolution copies. In Example 1 given above, while it is feasible to store approximately 420 hours of standard definition PAL video on the 10TB server, this would be as:

- 420 hours of a single copy, with no backup or recovery copies;
- 210 hours for the prime and recovery copy; or
- 140 hours for three copies.

However, the last two configurations would provide no physical separation between the copies.

At this stage data tape is the most cost effective storage solution for large collections of digital files. While magnetic tape is a well proven technology the formulation may vary from manufacturer to manufacturer. While the frequency of migration means that “life” is not required to be measured in decades as was the case with analogue media⁴ it may still be considered a risk to rely on a single manufacturer. Prior to accepting a batch of data tape for use samples should be tested to simulate the conditions under which the tape will be used and stored. While this carries a cost it is not as expensive as trying to recover lost data in the event of a problem.

Each of the three copies needs to be stored discretely. It makes little sense to have all three copies stored on a single tape or a single server. To spread the risk further each copy should be stored in as separate location, the further the geographic separation the better. Ideally each location should be subject to a different set of environmental or physical risks.

Example

Storing copies in different parts of the same building is more effective in reducing risk than storing both copies in a single room, however this practice is less effective in minimising the risk than separately storing the two copies in different buildings. The physical risk is further reduced if the two buildings have a great deal of geographical separation.

same room > different rooms > different building > different city

Reduction of environmental risks →

The decision on the file format chosen to store the data should involve consideration of the support for the format and especially if it is an open source or proprietary format. Proprietary formats may have features

⁴ The practice of relying on media life has issues of format obsolescence that have proven to be very difficult to manage, for example 2” video tape is very difficult to duplicate as playback equipment and skilled operators are very few, even globally.

that are desirable however the life of the format and the support is at the discretion of the format owner. Open source formats are frequently supported by a standard formulated by a community of experts. This standard describes the way the file is played and since the source is open there are more options in recovering the data should a problem occur. An open source format has a community supporting the format's development. Change is fed into the development from a wide range of perspectives and changes voted on by the community. This open communication offers a greater certainty and information on the timetable for migration.

Archives are about information, not only the information contained in the records but also about the records. This later information may be regarding the provenance of the records or technical details about the characteristics of the record itself. The expression *metadata*, or information about information, is used to describe these characteristics. In a digital collection metadata is crucial in managing the collection at every stage. In the past the technical information regarding the physical object was recorded in a database or appended to a catalogue. This practice could be continued for the files in a digital collection, however it would be far more difficult to manage the collection without having the metadata linked directly to the file. The importance of metadata is so great to a digital collection that should the metadata be lost or corrupted managing the collection of any size would be virtually impossible. For this reason there should be sufficient metadata embedded within the file structure so if the external copies of the metadata are corrupted or lost the file can still be identified and played back. There are two levels at which the metadata may be embedded:

- At the file level, many file formats have sections designed to store metadata. TIFF for still images, BWF for audio and AVI or MOV for video are examples of file formats that have been designed to carry metadata as well as content. MXF is a format designed to wrap coded audio and video along with extensive metadata to create a single file that has the potential to be exchanged between users.
- The second level is by “wrapping” groups of related files together in a single file package, such as TAR or Zip.

As has been mentioned a digital collection requires an active maintenance program. Some the maintenance may be automated within the server system, for example continuously, on-demand or “on-use” verifying of checksums to ensure that no corruption of the file has occurred. A major maintenance event is the migration of the data to new current hardware and formats on a regular or as needed basis to maintain the data in a readily accessible and supported format. The frequency of the migration is open to the changes in technological development, even open formats change. For example the Linear Tape Open (LTO), first released in 2000, is currently on the 5th generation with the 6th generation expected to be released later this year.

Any computer system, either storage or digitization, will lose data if there are no measures in place to secure a controlled shutdown in the event of loss of power supply. If you are making the one possible replay pass from a degraded carrier and your system goes down before the file is fully written, then you run the risk of losing the original and the duplicate file. Files must be properly written to be viable.

Instantly activated backup supplies or at least an Uninterruptible Power Supply (UPS) system to enable a controlled shut down and prevent loss of data are crucial steps to mitigate this risk. Additionally data storage systems should always operate in a copy—then verify—then delete method, so that a file is never in limbo if a system fails at a crucial point. SAN controllers often have internal battery supplies so that the controller will cache data in the event of a mains supply failure.

3. Recovery After a Disaster

The nature of the disaster will influence the approach taken to the recovery operation. If there has been hardware failure recovery may be as simple as recovering from an error on a RAID system,⁵ or recovery may involve complex conservation treatments to tapes or hardware and computer forensics. It is down to the successful development and implementation of the preservation strategy as to how complex, and expensive, the disaster recovery operation will be.

There is no “one size fits all” disaster recovery plan as each disaster will have differing direct effects and scale. There is a wealth of information on developing organisational disaster recovery plans available. In general terms the recovery plan should incorporate several distinct stages:

- *Evaluation of the impact of the disaster*

An information gathering step where the scope and scale of the disaster is assessed and the information feeds into the development of a plan of how to best to proceed and permits a draft budget to be formulated.

- *Salvage*

More relevant for physical objects that have been affected by a major disaster where the storage environment has been damaged to some extent by water, fire or collapse. For a digital collection this may involve salvaging data tapes or server equipment or locating files resident on a hard drive where the operating system has become corrupted.

- *Stabilisation*

Again this step is more directly relevant to physical objects. After a disaster chemical and biological factors will affect the stability of the object. Stabilisation in this instance could apply to data tapes or to preventing corrosion affecting hardware.

- *Specific recovery actions*

These are the specialist skills needed to make the carrier (tape or hard disk) able to be played and recovering the data. The recovered data is then checked and stored on new media or system.

Recovering data from a severely damaged hard disk is a highly specialised skill and falls into the area of forensics. If the drive is still able to be run there is software readily available that may be able to recover files from an otherwise inaccessible disk. However the recovery is a very slow process and many products are designed for consumer level operating systems and drives, not the larger enterprise level installations.

In a well planned and executed preservation strategy that has included contingencies for disaster recovery in the event of a disaster, the recovery would hopefully only require the selection of the best surviving copy of the files from the various recovery tapes. In practice this will involve checking each of the files checksums to ensure no corruption has occurred selecting the intact files from the source and reconnecting the damaged or lost media with the database links and metadata to which they belong. In addition to restoring any lost digital media, the collection metadata should also be updated to reflect that the media file had been restored from a new source, in the event that it later turns out that the media is incorrectly identified or corrupted. At worst this may mean ingesting these files onto a rebuilt server system. This action is equivalent in time and effort to a full migration.

⁵ Redundant Array of Independent Disks (RAID) systems use a variety of strategies including total redundancy to store the data across several layers of disks that operate as a single disk. However if a problem should occur with one of the disks in the array the information on that disk can be recovered and the faulty disk replaced.

Case Study: Full recovery from the failure of a RAID system

Recently a 48 Terabyte fibre channel connected hard disk storage system was installed in a video production and archiving environment. The video work area has three edit suites equipped with four workstations, evenly divided between Apple OS-X and Microsoft Windows XP environments. The disk system is not used as a permanent collection archive but hosts production work in progress and reusable stock footage. It is also used a drop box for work to and from the institutions MAM system

The 24 disk storage system is equipped with enterprise class 2 Terabyte drives and is protected to the RAID 5 level. (that is, one disk failure can be automatically corrected). In the event that a disk fails, the system administrator is alerted and a system rebuild automatically begins to recreate the array without the failed disk. 2 spare disk drives are held locally and may be hot swapped into the system by non-skilled staff.

A single disk drive in the array failed and the faulty disk was replaced with a known good spare. The array rebuild began and was scheduled to take more than 15 hours. *In some instances, rebuilding a disk array may take many days or even weeks depending on the size of the hard disks.* While the rebuild was underway, a second disk drive began to exhibit data errors. As a result, the parity system of the RAID 5 system was unable to simply recreate all the data without errors.

All the media files stored in the storage system appeared to be intact after the initial rebuild, but had random vision and sound errors which were only detectable by real-time replay. *A checksum would have proven this in a very simple manner, but production networks do not usually use checksums because of the overhead of generation and verification.* The system had alerted the administrators to the corruption of the volumes, but could not list the corruptions to file level.

The disk storage is fully supported by its manufacturer, and has a very sophisticated file system, designed to protect data even when multiple drives fail. Over a period of 3 – 4 days, the manufacturer's engineers were able to remotely access the system and to reconstruct 100% of the data without error by using their special and proprietary support tools. New replacement disk drives were provided under the existing warranty agreement and are available again for immediate use if required. The failed disks were shipped back to the manufacturer so that diagnosis of the failures could be undertaken.

Conclusions:

If you are using a disk based storage system for archival data, RAID redundancy measures are essential and can protect against data loss.

The bigger the disk drives in a system, the greater the rebuild time will be, and consequently the higher the probability that another disk drive will fail completely or partially during the rebuild period.

A RAID disk system should have the level of redundancy appropriate to the risk of loss of data for your application. RAID 5 has 1 redundant disk, RAID 6 has 2 redundant disks, etc.

Maintain support contracts with experts in your disk system. They can help when you have failures.

Backup your data if it is valuable and irreplaceable.

3.1 Rebuilding from Recovery Copies

There are simple calculations based on the published technical capabilities of tapes and drives that give a very approximate time taken to recover files. These are given in the table below.

Tape format	Capacity	Data transfer rate (maximum)	Theoretical transfer times hours (minutes)	Realistic times (66%) hours (minutes)
LTO3	400 GB	80 MB/s	1.42 (85)	2 (128)
LTO4	800 GB	120 MB/s	1.90 (114)	3 (171)
LTO5	1.5 TB	140 MB/s	3.12 (187)	5 (281)
LTO6	2.5 TB	160 MB/s	4.55 (273)	7 (410)

However it needs to be remembered that these are the maximum and relate only to data transfer rates. The actual times taken to transfer the data will be influenced by the entire network infrastructure. Anecdotally a figure of around 2/3 or 66% of the theoretical bandwidth is a more realistic data rate. Additionally the files will need to be checked and verified, this effectively doubles the time taken.

Recovering from catastrophic failure of a main database and collection storage system requires a carefully planned data recovery strategy. If the main collection database server has been lost, then there are steps required to restore this system before the data storage recovery can begin. A collection database and asset management system should be able to track the media it controls, where these files reside and the available redundant copies.

In a well designed system the main collection database is hosted in a Virtual environment (VM), and the most recent backup image of the virtual machine may be restored to any compatible VM environment within hours, not days. Strategically, this is a very solid solution. The VM images are snapshots of working computer systems and hardware emulation in software that are largely platform independent. They can be maintained simply and can run on compatible environments on almost any hardware.

The second stage involves rolling back of the database proper onto the restored server environment. A regular backup and storage of transaction logs can give virtually 100% certainty that the database can be restored to a state very close to or even identical to the onset of the disaster event.

Using a hypothetical collection database of 1,800,000 items in a complex data model may require approximately 200 Gigabytes on disk but may contain up to one Terabyte of data when all the transaction and backup logs are considered. Restoring this database to an exact replica environment will take one to two days to achieve but relies on a proper backup regime for both the main database and the transaction logs of work underway at the time as close as possible to the system failure. If the disaster recovery backups have been corrupted then this will extend the period required.

Restoration of an entire data collection requires several steps.

1. Identify which tape or tapes the required media files are on. If this is not possible then the process will be a bulk restoration from tape without any “intelligent” targeted reading of the data.
2. Restore from backup tape location to disk storage where integrity checks/verification are carried out. Checksum tests may take greater than real time relative to the replay duration of the media files, this will potentially require substantial periods of computer processing. Data drives may

fail particularly when a massive number of read write cycles are carried out over a short period. Heads may become clogged and require cleaning, etc. Disk systems connected as data library cache, and as working locations for intensive calculations like checksum need sufficient read/write speed and network bandwidth to support multiple simultaneous operations. Typically, large numbers of disk spindles connected in high-performance RAID configurations as Direct Attached Storage coupled with multiple network interfaces are recommended.

3. Read from data tape can approach the theoretical limits described in the table, but only when the system components are all designed to deliver optimum performance.
4. Write the restored data to new tapes using a method compatible with the database/MAM, this will enable the media location to be known.
5. Audit the MAM/database to make sure there are no missing items. If there are missing items, develop a methodology for finding where they might exist on other tapes (Copy2) or be prepared for some decisions regarding redubbing, obtaining a new copy from the original source or another source.

3.2 Physical recovery of data tapes

As the most cost effective storage media, data tape is the most probable media storage for a digital collection to be affected by a disaster. Most data tapes use a traditional binder style tape with magnetic particles (MP tape). Fortunately there has been a body of knowledge developed around the recovery of MP tapes. While it is strongly recommended to gain specific conservation advice before attempting any recovery treatment the basic procedure is to:

1. Remove any dirt from the outside of the tape cassette shell
2. Check for physical damage to the cassette shell
3. Open the cassette shell and clean any loose dirt lodged inside. Check the condition of the tape and the tape path.
4. It may be necessary to use a low relative humidity treatment⁶ to dehydrate the binder if the tape has been subjected to water or high relative humidity.
5. Reassemble the cassette
6. Clean the tape surface, this will require cleaning equipment specific to the format. If the tape is stretched or distorted during cleaning is highly probable that the data will not be able to be recovered.
7. If the tape appears to be in good condition and is thoroughly cleaned then it may be cautiously played on correctly adjusted equipment and the data transferred. There may only be one opportunity to do this depending on the condition of the tape.

Even after the best effort has been made to recover the data the disaster may have caused too much damage for simple playback. At this point it may still be feasible to recover some of the data by reconstructing the data bit by bit.

⁶ This is commonly referred to as “baking”, however this term is misleading and has led to the total loss of the tape/data due to the temperature being too high. As a rule of thumb the temperature should not exceed 50°C. Some references do suggest temperatures higher than 50°C to “speed” the process up, this is not recommended.

4. Conclusions

Prevention is the best option. Manage the risks with a well considered preservation strategy that incorporates effective disaster risk mitigation, rather than risk the costs or total loss that can so easily occur as a result of a disaster.

Metadata and Formats for Digitization and Digital Preservation

Data, Documents, and Memory

A Taxonomy of Sources in Relation to Digital Preservation and Authenticity Metadata

Joseph T. Tennis

University of Washington Information School

Abstract

This paper describes efforts to operationalize methods of metadata application to digital records. We summarize theoretical and applied research from the field and from the InterPARES research project to establish a grounding in the issues and suggestions for metadata assignment. Key to our understanding of metadata is 1) its relation to documentation (a more narrative form of description of records), and 2) its permanence in relation to the records themselves. This paper closes by presenting a draft taxonomy of sources useful to decision-making in relation to metadata assignment and documentation creation.

Author

Joseph T. Tennis is an Assistant Professor at the Information School of the University of Washington, a member of the Textual Studies faculty at UW, and an Associate Member of the Peter Wall Institute for Advanced Study at The University of British Columbia. He has been active in the InterPARES research project since 2005, and currently serves as an advisor and researcher on metadata issues. He holds a B.A. in Religious Studies, an M.L.S., an Sp.L.I.S. in Book History, and the Ph.D. in Information Science from the University of Washington. He works in classification theory, the versioning of classification schemes and thesauri (a.k.a. subject ontology), and the comparative discursive analysis of metadata creation and evaluation, including archival metadata, both contemporary and historical.

1. Introduction

Current archival theory has pointed to the role of both documentation and metadata as key to attesting to the authenticity of body of records. Upon closer inspection, we see that even if this is what theory prescribes, the complexity and variety of this kind of documentation and metadata seems *confusingly rich*. On the one hand we understand metadata to be machine and human readable assertions about resources, in general. In the archival context we constrain that meaning by saying that there are two kinds of metadata, and that those metadata are focused specifically on records and aggregations of records. The two types are intrinsic and extrinsic metadata—those that are permanently linked to the record and those that are not (Gilliland, 2008). Following the work carried out in the InterPARES research project, we find that the complexity in this case lies in asserting what metadata are required to persist along with the record over time (InterPARES). If it is intrinsic to the record, we might assume that it should persist. Those that are extrinsic pose a different methodological problem. How do we assess what stays and what is deleted? And though we assume intrinsic metadata should persist with the record, they are not without their complications. Advancing archival theory problematizes our ability to wed metadata inseparably with digital records, claiming that it is itself a study in addition and deletion, requiring particular methodological commitments. With both intrinsic and extrinsic metadata we must say what stays and what goes.

Documentation is another complex theoretical concern. The complexity lies with the relationship between documentation and digital recordkeeping in particular, and recordkeeping in general. Evolving theory related to authenticity and the conception of the *fonds* finds that the archivist must document

interventions made by the creator and preserver (e.g., MacNeil 2008 & Millar, 2002). This means that in constructing a coherent picture of a fonds or describing the provenance of a body of records, theory guides the archivist to document his or her decisions made. This documentation must also follow the record through various stages of preservation—persisting along with the records.

The question surfaces: how do we make operational the theory of metadata and documentation in digital records preservation systems, given this level of complexity? The first step that could be taken is to create a taxonomy of these *sources* of information so that we might use that categorization to manage that type over time. This paper takes the first step in that direction by proposing such a draft taxonomy. This taxonomy accounts for the characteristics of metadata and documentation mentioned above, and adds to it from InterPARES research and contemporary theory found in the literature. The result is a rubric that can be used to make decisions about how to design out systems that can keep authentic digital records.

2. Metadata vs. Documentation

In its most common definition, metadata is data about data. However, this definition is not adequate to distinguish metadata from documentation. And this is a distinction, in the context of archival metadata theory and practice that we want to establish and maintain. For our purposes, *general* metadata is human- and machine-readable assertions about a resource, where resource is the World Wide Web Consortium's (w3c) term. Resource to the w3c is anything with an identity. We have scoped resource in our context to be records, and our assertions are the various things that can be said about records for the purpose of authenticity, preservation, and retrieval. We have scoped metadata thusly based on InterPARES research (InterPARES). Specifically, we have drawn on the InterPARES Benchmark Requirements Supporting the Presumption of Authenticity of Electronic Records (InterPARES 1: Authenticity Task Force, 2005) the Baseline Requirements Supporting the Production of Authentic Copies of Electronic Records (InterPARES 1: Authenticity Task Force, 2005), and the Chain of Preservation (COP) model (Preston, 2009; Duranti and Preston, 2008).

Examples of metadata drawn from these sources are the names of persons concurring in the formation of the record. InterPARES has identified five persons that can be identified with the generation of a digital record: author, addressee, writer, originator, and creator. In each of these cases we can fill in the blank:

“The addressee of the document is *x*”

Documentation has been a concept closely associated with all stages of the lifecycle of records, but has become an even richer concept in the digital environment and in relation to contemporary discussions of metadata. In discussions of archival appraisal and description we have evolving theory of how documentation is required for the presumption of authenticity and to reveal the details of the agents and actions that helped create any given fonds.

In the context of archival description, we have accounts from MacNeil about the differences between metadata and archival description (1995). In this paper MacNeil distinguishes archival description from metadata by claiming the latter is the view from the ground, while archival description is like a view from an airplane. What description provides is an overview of the whole body, history, and scope of the body of records. Metadata, on the other hand, serve as the raw material for archival description. They are what the archivist uses to construct their bird's-eye-view.

Given this perspective on metadata, we can see that much of it should be discarded at the point of archival description. That is, once the body of records has crossed the threshold of preservation, bringing with it all attached metadata, the archivist paints a picture of the body of records from all available evidence and then discards the evidence, including much of the metadata. For example, before records move to the preserver, for recordkeeping purposes we might trace which records had been destroyed. In our Chain of Preservation Model (Preston, 2009), this is activity A3.4.3. However, once the body of records crosses the threshold of preservation it may not be the decision of the creator to keep this information, and in this case, it would not be part of the preserver's work to keep track of it. Of course, one can imagine the opposite, but we can take this as our example. In this case, the creator of the records only wants a statement of what is in the fonds, not what has been destroyed from the fonds. The metadata associated with this decision to destroy records is no longer kept after it is entrusted to the preserver.

The fonds, as a unit of analysis in archival work, has undergone some close inspection in the literature. Both MacNeil (2008) and Millar (2002) have reexamined our assumptions about the way we talk about, and hence document, the creator and their body of records. MacNeil's concern with authenticity, arrangement, and archival description drew her to research intentions of both authors of scholarly texts and archivists. Seeing through this comparison that:

the theory of final intentions is underpinned by a particular ideology concerning the nature of artistic creation, i.e., the author as solitary genius. The principle of original order, for its part, is underpinned by particular ideologies concerning the nature of historical inquiry. Lehmann's articulation of the Prussian principle of original order in the latter part of the nineteenth century, for example, resonated with the ideology of "scientific" history." (MacNeil, 2008, p. 13)

The upshot of this work is that we must document our own actions as archivists in representing creator's intentions. Thus, it is not by metadata alone that we can best represent the fonds, its history, and its accumulation into its current state. MacNeil calls the rearrangement of records by different custodians the records' archivality.

Millar in a not dissimilar vein, separates *respect des fonds* and provenance as two distinct concepts with which archivists must reckon. Her case study is the Hudson Bay Company's records. They are spread out over different archival institutions, and they serve as a lesson in provenance. This concept, provenance, would in Millar's formulation, encompass three distinct histories: 1) creator history, 2) records history (or recordkeeping history), and 3) custodial history. These would constitute what she sees as the only useful guiding principle, especially when compared to the unrealizable concept of the fonds. She wants archivists to work with a new concept she calls *respect de provenance*, and they would do that by writing out the three different histories.

Both Millar's histories and MacNeil's creator's intentions, and subsequent change of records arrangement by the chain of custodians, require something more than metadata can offer. They require documentation. We have found the same need for documentation in the InterPARES research project. In the process of drafting a metadata application profile that is consistent with diplomatic assumptions about records, in accordance with the findings of the Benchmark and Baseline Requirements (InterPARES 1: Authenticity Task Force, 2005), established by InterPARES 1, and based on the Chain of Preservation model (COP model) (Preston, 2009; Duranti and Preston, 2008), we found that metadata alone could not maintain presumption of authenticity in digital records systems through time.

3. The Chain of Preservation

The lifecycle of a body of records has been represented in ideal form in the Chain of Preservation model (COP model) (Preston, 2009; Duranti and Preston, 2008). Through this model we have begun to enumerate the metadata required for the presumption of authenticity (Tennis and Rogers, 2012). We call our metadata the IPAM, which stands for InterPARES Authenticity Metadata. There are a total of 428 assertions made about records and their context in the IPAM. We have categorized them into 12 categories. They are given below.

- AT – attachments: Signals those items attached to the record—indication of attachments is necessary for the integrity of the record.
- AU – authentication: Those elements that indicate the identity of the persons involved in the creation of the record.
- B – archival bond: Those elements that illuminate the connection of the record to other records to which it relates, and its context, whether it is preserved or destroyed.
- D – date: Points in time in the life cycle of the record(s) that need to be documented.
- DO – external documentation: Links to information that governs preservation, transfer, and access to the record(s) over time.
- F – form: The rules of representation that determine the appearance of an entity and convey its meaning.
- H – handling: Representation of the office or officer formally competent and/or responsible for carrying out the action to which the record(s) relates or for the matter to which the record(s) pertains.
- L – location: Indications of where the record(s) are stored, backed up, duplicated.
- P – persons: Identification of individuals or legally defined entities who are the subject of rights and duties and are recognized by the juridical system as capable of or having the potential for acting legally with regard to the record(s)
- R – rights and access: restrictions or privileges that apply to the record(s).
- S – subject: The action or matter to which the record(s) pertain.
- T – technology: The carrier(s) of the form and content of the record.

Of these, documentation (DO) rivals persons (P) as the most frequent assertion made. We have established at least 46 links to external documentation as required for the presumption of authenticity of digital records. For example in the context of records creation we need to indicate which records were transferred (DO0), whether the records were modified (DO1), and whether the records were backed up (DO3). To assert digital records transfer, modification, or backup, we need links to external documentation. The other documentation deals with corrections to records, updates to records, access to records, etc. They deal with the integrity of the records, the systems in which they are kept, and serve as an attestation of what kind of interventions effect the form and content of the records as they move from creation to preservation.

4. Taxonomy of Sources

If we build directly out of Millar, MacNeil, and InterPARES we can see a categorization of metadata and documentation surface. There are two categories of metadata and three categories of documentation. The two categories of metadata are Identity Metadata and Integrity Metadata. Identity metadata comprise:

Table 1. Table of Identity Metadata.

D00	the date of <i>document</i> creation
D01	chronological date (and possibly time) of compilation and capture;
F01	documentary form—that is, whether the document is a report, a letter, a contract, etc.; and
T01	digital presentation—that is, file format, wrapper, encoding, etc.
D02	chronological date (and possibly time) of transmission from the originator;
D03	chronological date (and possibly time) of receipt and capture;
F01	documentary form—that is, whether the document is a report, a letter, a contract, etc.; and
T01	digital presentation—that is, file format, wrapper, encoding, etc.
P02	- <i>author(s)</i> —that is, the physical or juridical person(s) responsible for issuing the document;
AU04	- <i>subscription</i> —that is, the name of the author or writer appearing at the bottom of the document; and
AU05	- <i>qualification of signature</i> —that is, the mention of the title, capacity and/or address of the person or persons signing the document;
AT01	- indication of any attachments—that is, mention of autonomous digital objects linked inextricably to the document.
P03	- <i>writer(s)</i> —that is, the physical person(s) or position(s) responsible for articulating the content of the document;
P04	- <i>addressee(s)</i> —that is, the physical or juridical person(s) for whom the document is intended;
P05	- the physical person(s), position(s) or office(s) responsible for the electronic account or technical environment where the document is generated and/or from which the document is transmitted;[1]
P06	- <i>receiver(s) or recipient(s)</i> —that is, the physical or juridical person(s) to whom the document may be copied or blind copied for information purposes;
S01	- name of the action or matter—that is, the subject line(s) and/or the title at the top of the document;
AU01	- indication of the presence of a digital signature;
AU02	- <i>corroboration</i> —that is, an explicit mention of the means used to validate the document;
AU03	- <i>attestation</i> —that is, the validation of the document by those who took part in the issuing of it, and by witnesses to the action or to the 'signing' of the document;
B01	classification code; and
B04	planned disposition (if not evident in the classification code).
B02	registration number.
P01	The physical or juridical person who makes, receives or accumulates records by reason of its mandate/mission, functions or activities and who generates the highest-level aggregation in which the records belong (that is, the fonds). Syn.: creator.
R01	indication of copyright or other intellectual rights;
H01	name of handling office (if not evident in the classification code);
H02	name of office of primary responsibility (if not evident in the classification code and records retention schedule);
R02	access restriction code (if not evident in the classification code);
R03	access privileges code (if not evident in the classification code);

B03	vital record code (if not evident in the classification code); and
AN01	priority of transmission; (urgent, etc.)
D04	transmission date, time and/or place;
SS01	actions taken;
D05	dates and times of further action or transmission; and
AT02	information on any attachments—that is, mention of autonomous items that were linked inextricably to the document prior to its transmission for the document to accomplish its purpose.
F02	draft or version number;
D06	archival or filing date—that is, the date on which a record is officially incorporated into the creator's records;
AT03	indication of any annotations[5] or new attachments (e.g., records profiles);
R02	access restriction code (if applicable and if not evident in the classification code)—that is, indication of the person, position or office authorized to read the record;
R03	access privileges code (if applicable and if not evident in the classification code)—that is, indication of the person, position or office authorized to annotate the record, delete it, or remove it from the system;
B03	vital record code (if applicable and if not evident in the classification code)—that is, indication of the degree of importance of the record to continue the activity for which it was created or the business of the person/office that created it;[6] and
B04	planned disposition (if not evident in the classification code)—for example, removal from the live system to storage outside the system, transfer to the care of a trusted custodian, or scheduled deletion.
B01	expression of archival bond (e.g., via classification code, file identifier, record item identifier, dossier identifier, etc.);
P01	name of the creator—that is, the name of the physical or juridical person in whose archival fonds the record exists;
R01	indication of copyright or other intellectual rights (if applicable);[2]
B05	indication, as applicable, of the existence and location of duplicate records, whether inside or outside the record-making or recordkeeping systems and, in instances where duplicate records exist, which is the authoritative copy—that is, the instantiation of a record that is considered by the creator to be its official record and is usually subject to procedural controls that are not required for other instantiations;[3]
H01	name of the handling office (if not evident in the classification code)—that is, the person or office using the record to carry out business;
H02	name of the office of primary responsibility (if not evident in the classification code or the records retention schedule)—that is, the office given the formal competence for maintaining the authoritative version or copy of records belonging to a given class within a classification scheme;[4]
T02	indication of any technical changes to the records—for example, change of encoding, wrapper or format, upgrading from one version to another of an application, or conversion of several linked digital components to one component only—by embedding directly in the record digital components that were previously only linked to the record, such as audio, video, graphic or text elements like fonts;

Identity Metadata are permanent and fixed to the records of the creator. The majority of the other metadata is Integrity metadata—that is it is metadata that accounts for the handling of records in digital systems from this point of creation through to the point of permanent preservation. By definition integrity metadata can be compiled as reports in external documentation. Thus, though the system may generate metadata, this metadata is then compiled into documentation and the metadata discarded as no longer necessary. Integrity metadata are further erased as they are reported in Creation, Recordkeeping, and Preservation Documentation.

The three categories of documentation follow the stages in the COP model. Creation documentation includes the transfer of records from the context of creation to the recordkeeping system. Recordkeeping Documentation is the outcome of MacNeil’s archivalterity, that is, the acts of continuous and discontinuous change that transform the meaning and authenticity of a fonds as it is transmitted over time and space (MacNeil, 2008 p.14). This kind of documentation also reflects the *custodial bond* “meaning the relations that exist between a body of records and the various custodial authorities that interact with the records over time, including archivists and archival institutions,” (MacNeil, 2008 p. 14). Any transfer, modification, correction, updates, refreshing that happens to records as they are kept is also reflected, in summary form, in this kind of documentation.

The final documentation is Preservation Documentation. We hypothesize that this category of documentation is relevant from a contingent definition of preserver. The trusted third party and ultimate keeper of the body of records. Ultimate here meaning *the current keeper considered the final keeper*. Once records move (and they do move) we move this documentation into to recordkeeping documentation. Preservation documentation consists of authentication reports, preservation feasibility reports, disposition reports, state-of-records reports (documenting technological carriers and documentary form of records as they cross the threshold of preservation). Preservation documentation also includes Millar’s creator history (documenting functional changes, and name the potentially diverse set persons involved in the creation of the fonds etc.), recordkeeping history (which would bring forward all the relevant integrity metadata), and custodial history (which would add to the transfer reports a narrative of context about where records were found, how and why they moved, and attempt to make clear the decisions of previous archivists).

5. Toward Operationalized Theory

To consider a taxonomy of sources is to take a step toward operationalizing theory. I have made a bold statement in outlining what metadata I think should be kept permanently and which can be transformed into documentation that follows the records. To tell the story of digital records is a complex task. What contemporary theory of archives tells us is that we must make clear our interventions, narrate our roles and how we see the roles and actions of others. This supplements our understanding of archival description, makes clear the role of temporary and permanent metadata, and makes robust our systems of memory.

References

Duranti, L. and R. Preston, eds. *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive, and Dynamic Records* Padova: Associazione

- Nazionale Archivistica Italiana, 2008. Online reprint available at <http://www.interpares.org/ip2/book.cfm>.
- Gilliand, A. "Setting the stage." In *Introduction to Metadata*, edited by M. Baca, 1-19. Los Angeles: Getty Publications, 2008. Online edition available at http://www.getty.edu/research/publications/electronic_publications/intrometadata/index.html.
- InterPARES. www.interpares.org.
- InterPARES 1: Authenticity Task Force. "Appendix 2: Requirements for Assessing and Maintaining the Authenticity of Electronic Records." In *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, edited by Luciana Duranti, 204-219. San Miniato, Italy: Archilab, 2005. Online reprint available at http://www.interpares.org/book/interpares_book_k_app02.pdf.
- MacNeil, H. "Archivalterity: Rethinking Original Order." *Achivaria* 66 (2008): 1-24.
- MacNeil, H. "Metadata strategies and archival description: comparing apples and oranges." *Archivaria* 39 (1995): 22-32.
- Millar, L. "The Death of the Fonds and the Resurrection of Provenance: Archival Context in Space and Time." *Archivaria* 53 (2002): 1-15.
- Preston, Randy. "InterPARES 2 Chain of Preservation (COP) Model Metadata (Draft)." 2009. http://www.interpares.org/rws/display_file.cfm?doc=IP2-cop-model_metadata_v1.0.doc.
- Tennis, J. T., and C. Rogers. "Authenticity Metadata and the IPAM: Progress toward the InterPARES Application Profile." In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2012*, 38-45. Kuching/Sarawak, Malaysia, 2012. <http://dcevents.dublincore.org/IntConf/dc-2012/paper/view/109>.
- UNESCO. *Universal Declaration on Cultural Diversity*. 2001. Accessed 1 February 2009. <http://www.un-documents.net/udcd.htm>.
- UNESCO. *UNESCO World Report: Investing in Cultural Diversity and Intercultural Dialogue*. Paris: UNESCO Publishing, 2009.
- Wikipedia. "Optical character recognition." Accessed 1 September 2012). http://en.wikipedia.org/wiki/Optical_character_recognition.

Ensuring a Future for the Past

Long-term Preservation Strategies for Digital Archaeological Data

Adam Rabinowitz,¹ Maria Esteva² and Jessica Trelogan¹

The University of Texas at Austin, ¹Institute of Classical Archaeology and ²Texas Advanced Computing Center

Abstract

As the documentation of archaeological research is increasingly born digital, the preservation of archaeological knowledge is more and more dependent on the documentation and long-term curation of those digital files themselves. Because archaeological investigation is destructive, excavation records provide the only source of evidence for contextual relationships. Digital archaeological records are threatened not only by technical change and equipment failure, but by insufficient metadata. This paper describes the collaborative efforts of the Institute of Classical Archaeology and the Texas Advanced Computing Center to develop strategies to ensure the preservation and accessibility of the digital archaeological record in the long term.

Authors

Adam Rabinowitz is Assistant Professor of Classics and Assistant Director of the Institute of Classical Archaeology at the University of Texas at Austin. He holds a Ph.D. (2004) from the Interdepartmental Program in Classical Art and Archaeology at the University of Michigan and is a 2002 Fellow of the American Academy in Rome. He has more than twenty years of archaeological experience at Greek, Roman, and Byzantine sites in Italy, England, Israel, Tunisia, and Ukraine. His current projects include the publication of recent excavations at the site of Chersonesos (Ukraine), which will involve an online, GIS-enabled database of primary documentation.

Maria Esteva has been the Data Curator at the Texas Advanced Computing Center in the Data and Collections Management Group since receiving her Ph.D. in Information Science from the School of Information at the University of Texas at Austin in 2008. Her work includes the design of metadata architectures for evolving scientific and humanities data collections, the implementation of data collections management and archiving workflows, and the training of researchers in diverse aspects of data management. Esteva is the PI of an ongoing research collaboration with the National Archives and Records Administration (NARA) that investigates the use of computational and visual analytics methods for archival processing. She has presented her work at the Digital Humanities, Digital Curation, ASIST, IS &T Archiving, SAA and VIS/VAST conferences.

Jessica Trelogan is a Research Associate at ICA specializing in GIS and remote sensing technologies for archaeological applications. She has an MA in Classics (1994), and has been involved in all stages of archaeological research—from the field to publication to digital archiving—at ICA's excavation, conservation, and survey projects in Italy and Ukraine.

1. The nature of the problem

As the UNESCO Memory of the World program recognizes, digital tools have exponentially expanded our ability to create and share documents of all forms. At the same time, this vast collection of digital data is disturbingly fragile, as digital files and systems themselves require active curation if they are to survive

in a usable form even a few decades into the future. The tension between the possibilities and risks of digital documentation is nowhere more clear than in the field of archaeology.

In most humanities research, digitization or digital recording provides an additional layer of documentation for texts or objects that exist in the physical world. If that layer of digital information disappears, the physical items it represented can still be examined. The use of digital documentation strategies, then, only offers advantages: if digital versions are preserved, they provide an additional mechanism for the long-term survival of the information contained in the item, and if they are not preserved, the sum of knowledge is at least no less than it was before the digital version was created. This fact has been recognized by those in charge of archaeological archives consisting of paper documents and film negatives, and both major and minor projects focused on the digitization of such archives have emerged in recent years.¹ The challenges in these projects have been mainly related to the modeling of the data collections and the development of methods to share them effectively. Digitization in this case is a low-risk, high-reward strategy for preservation and dissemination.

The situation is quite different, however, for the use of digital records in the primary documentation of current archaeological research, especially when that research involves excavation. The excavation of an archaeological site, by its very nature, destroys the object of investigation. After a dig, the contextual relationships between artefacts, soil deposits, buildings, and other elements of the human environment only survive in the project's documentation. Perhaps even more than other humanistic disciplines, archaeology has been quick to adopt sophisticated digital tools, and the use of everything from digital photography to computer databases to three-dimensional modeling is now widespread in the field. As a result, the contextual record for many archaeological sites now exists primarily in digital form.

On one level, this has been a boon: digital tools have led to unprecedented levels of detail in field recording, better means of visualization of the process of excavation, and a dramatically increased ability to share digital and digitized data in all their original contextual splendor. Digital photography has freed projects from the financial constraints of film and development costs, and excavations that were once documented with a few hundred photographs are now documented with thousands, if not tens of thousands, of digital images. Digital frameworks composed of relational databases and Geographic Information Systems (GIS) allow information to be queried and filtered in ways that promote new understanding and interpretations. Moreover, the integration of digital datasets from different projects allows new questions to be asked and inspires new observations, and web-based interfaces have opened the results of research to the public on a global level.

In order for data to be shared and preserved, however, they must also be documented in ways that allow them to be compared and reconstructed in the future. Over the last 15 years, several centralized repositories for archaeological data have emerged, with their primary task the creation of metadata and

¹ One of the most visible of these efforts in the field of Classical Archaeology is the digitization of the archives of the excavations carried out by the American School of Classical Studies at Athens in the Athenian Agora and at the site of ancient Corinth ("Digital Collections, The American School of Classical Studies at Athens," accessed September 14, 2012, <http://www.asca.net/research?v=default>). Two of the authors of this paper have been involved in such digitization projects both at our home institution, the Institute of Classical Archaeology, and at the site of Chersonesos in Crimea, where substantial portions of more than a century of archival records have been digitized and presented online ("K. K. Kostsyushko-Valyuzhinich and his Reports for the Imperial Archaeological Commission," accessed September 14, 2012, <http://kostsyushko.chersonesos.org/>; "Discovering Chersonesos," accessed September 14, 2012, <http://www.discovering.chersonesos.org>).

the curation and migration of original datasets.² At the same time, various groups have worked on metadata ontologies and schemata that allow highly heterogeneous archaeological data to be described in ways that are mutually comprehensible between archaeological communities and across national boundaries.³ These efforts in the archaeological sphere are similar to efforts in other spheres of digital documentation, which, guided to a large extent by library practice, have focused on metadata standards and formal state-level or institutional digital repositories.

Digital archaeological documentation presents much greater barriers to standardization, however, due to its heterogeneity, its level of relational complexity, its use of a wide variety of proprietary or custom-built software platforms to manage contextual relationships, and the extremely varied and idiosyncratic circumstances of its production.⁴ The general archival schemata most frequently used by libraries are ill-fitted to the contextual relationships embedded in archaeological documentation, while custom-built schemata or ontologies designed to reflect the full complexity of archaeological data are inevitably so complicated themselves that they are very difficult for either non-information-scientists or non-domain-experts to understand. The varied environments in which archaeological data are produced also work against centralizing efforts. Where there is a relatively small, homogeneous community of practice and a strong incentive (e.g., a legal requirement) for the submission of digital data to a repository, centralization has been an effective strategy: in the UK, for example, the ADS, by governmental mandate, is the repository of record for all information produced by contract archaeologists. Where the community is large, diverse, fractious, and lacking in incentives, on the other hand, centralization tends to fail. In Mediterranean archaeology, projects are still more likely to request funds to build their own unique databases, in which they use their own idiosyncratic terminologies and metadata structures, than they are to request funds to allow them to deposit their data in an existing repository.

The issue of decentralization is compounded by national and cultural factors: not only do practitioners in different countries use different languages and describe and organize their material differently, but many archaeologists are still intensely suspicious of digital repositories and deeply reluctant to share the results of their research in any form other than traditional paper publications. Since the professional incentives for most academic archaeologists still center on publications “branded” by individuals or at least by individual projects, even those scholars who embrace the potential of digital media largely focus on short-term goals like attractive, customized websites rather than less exciting issues of long-term preservation.

² The most prominent of these are the UK’s Archaeology Data Service (“Archaeology Data Service Homepage,” accessed September 14, 2012, <http://archaeologydataservice.ac.uk/>), established in 1996, and the US-based Digital Archaeological Record (“tDAR,” accessed September 14, 2012, <http://www.tdar.org/>).

³ These efforts have taken several forms, one of which focuses on the creation of semantic frameworks and formal ontologies for cultural heritage material (e.g., the CIDOC-CRM: “The CIDOC Conceptual Reference Model,” accessed September 14, 2012, <http://www.cidoc-crm.org/>), another of which combines a domain-specific ontology with a specific metadata schema (e.g., OCHRE and ArchaeoML: “Online Cultural Heritage Research Environment,” accessed September 14, 2012, <http://ochre.lib.uchicago.edu/>), and yet another of which focuses on the adaptation of existing metadata standards like Dublin Core.

⁴ In the field of Mediterranean archaeology alone, one finds everything from small-scale academic projects run by one individual to large-scale academic projects carried out by several institutions, long-term excavation projects with 100-year histories controlled by one country’s research bases in another country, contract archaeology carried out by for-profit companies, rescue archaeology carried out by the staff of national archaeological services, museum research, etc.

This introduces the dark side of digital documentation in archaeology: few or no provisions have been made for the long-term preservation of, and access to, the vast majority of digital archaeological data. Many, if not most, digital archaeological collections are at risk on two major levels, even if we leave aside the question of the longevity of storage media. The first threat lies in the obsolescence of the proprietary applications that manage contextual relationships. Most archaeologists who were already working in the field during the advent of digital methods wince when they remember databases in DBIII that are now unrecoverable. The second threat lies in insufficient metadata for individual files associated with complex digital records, either because the intermediary program that managed the metadata is no longer accessible or because metadata was never created for these files in the first place.⁵ In the latter case, the metadata exists only in the heads of the excavators, and without access to those personal memories, much, if not all, of the context surrounding a given piece of documentation is lost. The loss of that context means the loss of some or all of the information contained in that file—and when that file documented a feature that was destroyed in the course of archaeological investigation, this means the permanent and irrevocable loss of that part of the memory of the world.

This is no longer primarily a technical problem for archaeology, for many of the technical issues have already been addressed by existing domain-specific initiatives. The archiving community has established standards and best practices for the curation and migration of many types of digital files; the community of information scientists associated with archaeology and cultural heritage has created ontologies and metadata schemata that are suitable for almost all the types of information archaeologists currently produce; and several effective, sustainable repository infrastructures are in place.⁶ It continues, however, to be a human problem. The creation of extensive metadata and the curation of files and formats is time-consuming and unrewarded, and therefore a low priority. Metadata standards are hard to understand and harder to apply to an existing dataset without the help of a specialist in information or library science. The submission of a dataset to a repository usually involves both the loss of control and diminished functionality, and requires that the dataset remain static and unchanging, all drawbacks that are anathema to many directors of archaeological projects. And all of these things cost money that few archaeological projects, especially outside the first world, can spare.

To escape the cloud of digital amnesia now looming over the field of archaeology, we must address not only the technical side of the question, but the human side as well. It is crucial to recognize the barriers—cultural, psychological, financial, and disciplinary—that prevent the creators of archaeological data from taking effective measures for their long-term preservation. Any solution has to involve not only centralized repositories and universal standards, but also tools that will work for individual data owners on a cellular and distributed level. Such tools should make it fast, easy, and cheap for those data owners to provide metadata for their collections. They should bridge the knowledge gap that deters many domain specialists from dealing with metadata standards and archival practices. They should allow data owners to preserve the idiosyncratic conceptual and organizational principles that structure their datasets. Finally, they should offer clear short-term rewards, either making it easier to share and publish data online or lowering the costs in time and money of the eventual transfer of a dataset to a repository.

⁵ Examples are digital photographs taken of a stratigraphic layer or find during excavation, which in many databases are embedded or referred to without being given metadata of their own on a file level, or the individual shapefiles or feature classes in a GIS, which are difficult to understand without additional documentation.

⁶ See, for example, the extensive guides to best practice prepared by ADS and tDAR: “Archaeology Data Service/Digital Antiquity Guides to Good Practice,” accessed September 14, 2012, <http://guides.archaeologydataservice.ac.uk/g2gp/Contents>.

The collaboration between the Institute of Classical Archaeology (ICA) and the Texas Advanced Computing Center (TACC) at The University of Texas at Austin, now in its fourth year, has been focused on these issues. It developed because ICA found itself in exactly the sort of situation described above: it possessed a large quantity of digitized archaeological data and a growing set of born-digital data, much of which was in proprietary formats or managed by proprietary applications, and much of which lacked metadata on the file level that could, in the absence of a database or the excavator, explain what was represented, for example, in a digital image. Together with this increasingly unmanageable digital dataset, a series of equipment failures and file corruptions provided compelling reasons to seek a long-term solution. The dataset was too large, too complex, and too much in flux for it to be handled by the digital repository of the UT library system, however, and the centralized archaeological repositories that existed at the time could not offer to preserve the full database functionality and spatial tools that ICA saw as essential to the management and publication of its data.

TACC, on the other hand, was very familiar with the management of complex, dynamic datasets, and its services are available at low or no cost to UT research projects.⁷ Corral, its data facility, had the technological infrastructure to support the full functionality of ICA's existing data management solutions: this resource consists of six petabytes of online disk space and supports databases, web-based access, a high-performance parallel file system, and other network protocols for storage and retrieval of data. For database collections, Corral provides DB server nodes running MySQL, PostgreSQL, and SQL Server, as well as support for open source domain specific databases; a GIS server to allow web and desktop access to spatial datasets is also being implemented. Most importantly, TACC had recently deployed a new Data and Collections Management Group (DCM) to support data intensive research activities. The DCM group designs, builds, and maintains the data applications facilities and consults with researchers in aspects of their collections lifecycle, from creation to long-term preservation and access. Group members are specialized in software development, Relational Database Management Systems (RDBMS), Geographical Information Systems (GIS), scientific data formats, metadata, large storage architecture, systems administration, digital archiving and long-term preservation. At the same time, TACC was also seeking collaborations in the digital humanities, and thus a joint project was mutually beneficial.

2. The initial ICA-TACC collaboration: excavation data from Chersonesos, Ukraine

2.1. The dataset

This project began with a primarily born-digital dataset created in the course of excavations at the Greek, Roman, and Byzantine site of Chersonesos in Crimea, Ukraine. These excavations, carried out between 2001 and 2006, had focused on a residential block in the urban center of the site with a continuous 2000-year record of occupation and a wide range of archaeological material, from ceramics and metal objects to human remains, charred seeds and metalworking waste. Contextual records were kept in a digital database, first in Microsoft Access®, then in a SQL database with an Access front-end, and then in ARK

⁷ Maria Esteva, Christopher Jordan, Tomislav Urban, and David Walling, "Cyberinfrastructure supporting evolving data collections," in *Proceedings of the 8th International Conference on Preservation of Digital Objects, iPRES 2011, Singapore, November 1-4, 2011*, ed. J. Borbinha Borbinha, A. Jatowt, S. Foo, S. Sugimoto, C. Khoo, and R. Buddharaju (Singapore: National Library of Singapore and Nanyang Technical University, 2011), 93-96.

(the Archaeological Recording Kit), a SQL database with a web-based front end.⁸ This database also managed a large number of digital photographs of excavation and objects, and it was linked to a GIS that included two- and three-dimensional data in both vector and raster form. Other born-digital data from Chersonesos included specialist spreadsheets, 3D models of objects, and a set of interactively-lighted image files (reflectance transformation images), each of which was derived through the processing of a large number of related individual digital photographs.⁹ In addition to these born-digital records, the digital dataset also contained digitized versions of paper recording sheets, hand-drawn plans, and hand-linked object illustrations, as well as both scans and electronic transcriptions of excavation notebooks.

ARK serves as a tool for the management and presentation of archaeological data, but it is not intended as a long-term preservation system, nor does it provide a complete representation of all the objects produced through the research process. Furthermore, some files in the dataset were in proprietary formats, while even files in open formats were often managed by proprietary systems such as ESRI's ArcGIS. The problem of organizing and preserving the original raw data, as well as the data produced off-site and after field seasons, still had to be resolved. Increasing numbers of DVDs, hard-drives, two servers, and personal computers filled with unlabeled or inconsistently labeled data were overwhelming the researchers and undermining the potential for the future reuse of those data.

2.2. Evaluation and triage

During initial conversations and interviews between the DCM group and the ICA team, it became clear that the project had complex requirements. ICA needed to archive data and it also needed an integrated digital environment in which different team members could organize and keep track of active and archival data within a dataset that was both changing and growing as research progressed. As a first step, the team carried out a triage of the Chersonesos collection. This triage was an extensive preliminary examination of the collection that aimed to define its structure, identify the types and locations of its various components, and describe the research workflows that produce those components. To approach this task, we used Records Management concepts and practices, including functional analysis and collections inventorying.¹⁰ Our analysis was able to identify four broad research functions with associated data types and workflows: on-site data collection in the course of excavation, post-collection processing of both excavated objects and digital information (both during and after the field season), analysis and interpretation, and publication. We then designed a system to inventory groups of objects in relation to these research functions. For practical reasons, this system focused not on the documentation of individual items, but on the description of groups of digital objects, for which type, location, and relations

⁸ For a description of the evolution of the recording system, see Adam Rabinowitz, Stuart Eve, and Jessica Trelogan. "Precision, accuracy, and the fate of the data: experiments in site recording at Chersonesos, Ukraine," in *Digital Discovery: Exploring New Frontiers in Human Heritage, CAA 2006*, ed. Jeffrey Clark and Emily Hagemester (Budapest: Archaeolingua, 2007), 243-256. For discussion of ARK, see Stuart Eve and Guy Hunt, "ARK: a developmental framework for archaeological recording," in *Layers of Perception: Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, Germany, April 2-6, 2007*, ed. Axel Poluschny and Karsten Lambers (Bonn: Deutsches Archäologisches Institut; Dr. Rudolph Hablet GmbH, 2008).

⁹ Adam Rabinowitz, Carla Schroer, and Mark Mudge, "Grass-roots imaging: a case study in sustainable heritage documentation at Chersonesos, Ukraine," in *Making History Interactive: Proceedings of CAA 2009*, ed. B. Frischer and L. Fisher (Budapest: Archaeolingua, 2010).

¹⁰ William Saffady, *Managing Electronic Records*, 4th ed. (Lenexa, KS:ARMA International, 2009).

to other groups of objects or records—such as “component of”, “derivative of”, “description of”, “complementary documentation of”—were defined.

This initial evaluation and description of the collection allowed both TACC and ICA collaborators to understand it more thoroughly.¹¹ The process revealed the variety of objects included in the collection and highlighted the need to impose a more logical structure to identify systematically their types, roles, and content, but it was also limited and inefficient. It required the efforts of TACC information scientists, a graduate student from the UT School of Information, and several undergraduate students, in addition to those of ICA and TACC staff members, and the inventory and organization process alone took nearly two years. Even after this process, a substantial number of digital objects lacked full identifying information. Attaching metadata by hand to all of the files would have involved prohibitive costs in time and person-hours. We also wanted to be able to apply the large quantity of information about data objects and their contextual relationships that was already encoded in the ARK database. The answer lay in the creation of a semi-automated solution for the generation of object-level metadata.

2.3. Metadata extraction

The solution centered on the use of iRODS,¹² a rules-based storage infrastructure deployed on Corral. iRODS acts as a management platform for archival collections, while the preservation of the data it contains is ensured by replication at TACC’s Ranch tape archive and at other HPC sites in Texas and across the country. Within such an environment, research teams can seamlessly integrate data management activities throughout the various stages of research. In the case of the Chersonesos dataset, our study of the collection and its workflows led to a decision to implement a file-management strategy that would also produce metadata automatically as data were ingested into iRODS for storage.¹³ This strategy involves two basic components: a standardized file naming convention, and a hierarchically labeled directory structure to classify and group related data. Both components were designed to provide descriptive, structural and contextual metadata that can be parsed and mapped to standards and that reflect the project’s data lifecycle. An ingest script on the iRODS side uses the naming convention and the directory hierarchy to extract such contextual metadata, together with preservation metadata, and encode them as XML Dublin Core (DC), Preservation Metadata Maintenance Activity (PREMIS), and Metadata Encoding and Transmission Schema (METS) metadata schemata.¹⁴

The file naming convention contains four information elements: a) an alphabetic code that indicates the ARK database module to which the digital object should be attached; b) the alphanumeric code assigned in the field and in the ARK database to the object, context, or record represented by the data object; c) the stage of the research process in which the data object was produced; and d) the designation of the file as either a master (an unmodified original) or a version (a derivative of a master file). The research stage

¹¹ Maria Esteva, Jessica Trelogan, Adam Rabinowitz, David Walling, and Stephen Pipkin, “From the site to long-term preservation: a reflexive system to manage and archive digital archaeological data,” in *Archiving 2010. Proceedings of the Archiving Conference*, vol. 7 (Society for Imaging Science and Technology, 2010), 1-6.

¹² “iRODS Data Grids, Digital Libraries, Persistent Archives and Real-time Data Systems,” accessed September 14, 2012, https://www.irods.org/index.php/IRODS:Data_Grids,_Digital_Libraries,_Persistent_Archives,_and_Real-time_Data_Systems.

¹³ David Walling and Maria Esteva, “Automating the extraction of metadata from archaeological data using iRods rules,” *International Journal of Digital Curation* 6, no. 2 (2011), accessed September 14, 2012, doi:10.2218/ijdc.v6i2.20.

¹⁴ “Library of Congress Metadata Schemas,” accessed September 14, 2012, <http://www.loc.gov/standards>.

designations were developed by the researchers; they apply primarily to images of objects and include: ‘b’=before conservation, ‘d’=during conservation, ‘a’=after conservation, ‘l’=lifting, ‘m’=microscope, and ‘s’=studio. Multiple data objects may be produced for a given archaeological object at a given research stage, so a sequence is recorded numerically following the stage designation. For example, for the data object (in this case, an image) named ‘sfi_CH05SR_3065_a1_m.JPG’, ‘sfi’ identifies the file as associated with the “special finds” module in ARK; ‘CH05SR_3065’ identifies the object represented as registry number 3065 from Chersonesos excavations in the city’s South Region in 2005 and matches the item key for this object in ARK, ‘a1’ indicates that the file is the first of several created after the object’s conservation, and ‘m’ indicates that it is the original version of the file (in this case, the original jpg generated by the camera when the photo was taken). Naming an object in accordance with this convention allows for the automatic extraction of metadata about that object, both from the stage codes embedded directly in the file name and from an external data management system like ARK through the inclusion of item keys that match records in the database. In the latter case, the matching of the item key also allows the automatic extraction of related metadata, so that, for example, information about the stratigraphic context of the find can also be included in the metadata record for the image file.

The hierarchical directory structure (**Figure 1**) serves to categorize the data as it is gathered and produced during the different research stages. Top-level directories are labeled according to documentation type. For each type, the sub-directories within reflect the materials to which that type of documentation is applied, and then the different kinds of documentation that are generated during the analysis, interpretation and publication of these materials. When a given data object is placed in a particular folder and accessioned into iRODS, metadata reflecting all the other classifications implicit in the directory path leading to that folder is extracted and mapped automatically to the DC “subject” element.

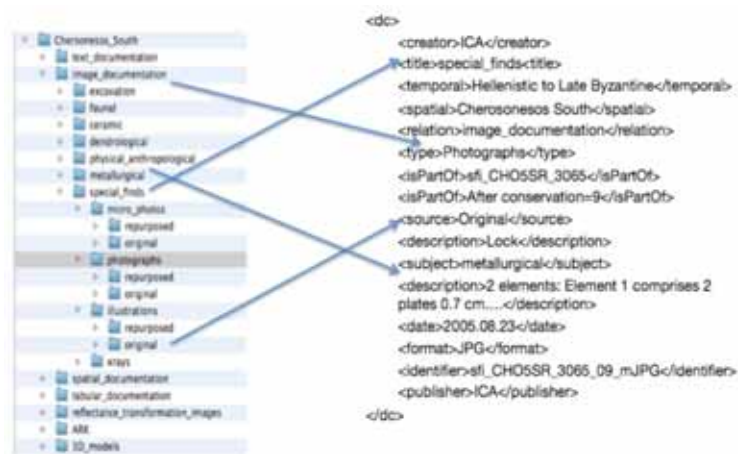


Figure 1. Hierarchical directory structure devised to classify ICA data.

Technical metadata from the files is extracted using FITS and mapped to PREMIS. A METS document is generated to contain all the metadata schemata. As a result, each object stored in the recordkeeping system on iRODS has a METS metadata record. The descriptive metadata is also registered on the iRODS metadata catalogue to enable search and retrieval through different available iRODS interfaces. By virtue of the object code, and the relationships established by the “isPartOf” element in DC, the resultant metadata acts as a glue that maintains the relationships between files and the components of the

archaeological record and tracks changes to these files across different research stages. Most importantly, these metadata records ensure that the contextual relationships between records, and records and files, are not dependent on the survival of the original database for long-term preservation.

The implementation of the processes for metadata extraction and mapping involved the creation of the metadata extractor script and the integration of this script into the iRODS rule engine, which manages the entire process. The workflow is exemplified in **Figure 2**. As objects are ingested to the collection, a Jython script is called to manage numerous sub-tasks for metadata manipulation. In turn, other rules enable controlled manipulation of data objects including removal, renaming, or relocation in the context of active collection management or in cases of misclassification of a file, while at the same time enforcing administrative approval for certain tasks. As a whole, the system constitutes a semi-automated platform that allows both the active management of the data objects in the collection and the generation of rich, contextual, continually updated metadata for each object.¹⁵

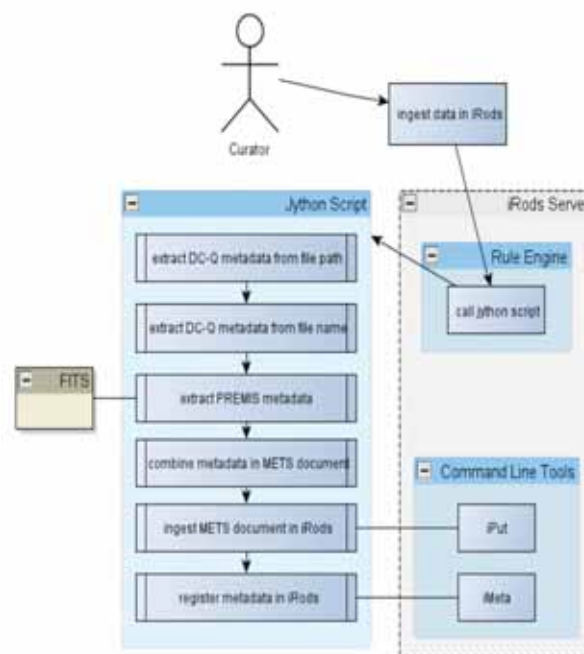


Figure 2. Workflow for automated metadata extraction using iRODS rules.

2.4. Problems and unresolved issues

After two years of using and testing this system, we have been able to evaluate its benefits and identify what needs to be improved from both human and systems perspectives. While the record-keeping system still requires archaeologists to manually name files and place them appropriately within the directory hierarchy, it automates the rest of the metadata generation process and provides, in conjunction with information retrieved from an external database, more extensive descriptive documentation for each file. Since naming and classification are manual steps, however, the extraction process must account for human error. Inconsistencies in the application of a file naming convention will cause the ingest script,

¹⁵ Walling and Esteva, “Automating the extraction of metadata.”

which can only document relationships on the basis of perfectly-formatted file names, to omit descriptive DC metadata. Poorly named files will thus be provided with a METS document containing only a PREMIS record. Furthermore, the system has a limited capacity to deal with files that lack 1:1 relationships with records in the database. For example, an image of a group of pottery fragments from several different stratigraphic contexts can only be associated by its file name with one of those contexts. Also, because the metadata is constructed automatically, the script must be manually customized to include collection-level metadata that applies to all digital objects, like project title, creator, or spatial and temporal coverage. All these issues create some barriers for the general application of this strategy.

3. Toward more broadly applicable solutions: excavation and survey data from Metaponto, Italy

3.1. The challenge of a new, more heterogeneous dataset

Our team spent a large amount of time designing a recordkeeping system tailored specifically to one excavation project, but the Chersonesos dataset represents only a small fraction of ICA's total digital collection. In fact, the ICA archives contain multiple datasets representing a series of projects carried out over more than three decades of research, each with its own characteristics and range of diverse data types. The bulk of ICA's collection is composed of information produced in the course of both excavation and archaeological survey in South Italy, especially in the rural territory of the ancient Greek site of Metapontion (modern Metaponto), from the early 1970s to the present. In the last ten years, many of the original analogue records produced by these projects, including maps and plans, negatives, slides, and paper recording sheets, have been digitized on an expedient and non-systematic basis. As a result, a large proportion of this analogue documentation now also exists in a variety of digital formats, from GIS shapefiles to scanned image files to word-processing documents. A significant number of these data lack sufficient metadata and organization, making it extremely difficult for researchers to work efficiently with them and almost impossible to share them with the larger public. While several small databases have been created by various specialists or eager students over the years to manage specific subsets of this material, especially images, most of the files are not associated with a database of any kind and almost none of them are documented with reference to their original provenance in the same detail as the Chersonesos data in ARK. While the Chersonesos excavation and documentation methodology relied explicitly on the individual stratigraphic context as its central relational unit, ICA's earlier excavations were organized around much more vaguely defined excavation areas. Furthermore, in the case of its survey datasets, which reflect entire ancient landscapes as opposed to structures, the central unit is the "archaeological site". This very different body of digitized data provided an excellent test-case for the expansion of the metadata-extraction strategies we deployed for the Chersonesos collection to datasets with different, looser structures and less extensive documentation.

The challenges to the integration of metadata from the different systems in which the Chersonesos dataset had been created and stored throughout the research process were, as we suspected, exponentially greater for the Metaponto datasets. Our solution for Chersonesos involved a high degree of automation, but it also required rigid and consistent rules for naming and for classifying the digital objects. This has made it very difficult to replicate the solution for the Metaponto collection, much of which was generated over a long period of time prior to the involvement of the current ICA research team and therefore presents little consistency in naming or classification. It is likely that a rules-based automated solution

would pose even more significant barriers to adoption by archaeological projects at other institutions, which would need not only to develop and adhere to their own strict naming conventions, but also to be able to implement an iRODS storage infrastructure and customize ingest scripts to match their own conventions and collection details. All of this would require a team of highly specialized software developers and metadata specialists that most institutions are unlikely to have in place.

3.2 Visualization-based strategies for triage

Our current goal, then, is to explore methods to capture standardized and complete metadata from heterogeneous and idiosyncratic archaeological collections without having to impose on them a rigid top-down system, and without depending on the ingest of datasets into a particular storage infrastructure. The first step in this exploration has been the identification of better methods for the initial triage of complex collections. As we discussed above, inventorying and organizing the Chersonesos collection took a significant amount of time and effort, even though the team was already intimately familiar with its contents, having been involved in the data production and curation from the beginning of the project. To process larger, more disorganized, and less familiar collections at ICA more efficiently, therefore, we are using and extending a visualization tool developed through a National Archives and Records Administration (NARA) research collaboration.¹⁶

The visualization application was developed for purposes of helping curators explore large and heterogeneous datasets with which they are not necessarily familiar. The tool allows for the visual exploration of a collection's structure according to several different organizing principles: its directories and sub-directories, its main data types, how many files are duplicated or contain errors and where these are located. It also makes it possible to visualize directory labels and file names as tag clouds, so that the user can quickly identify the most frequent names or labels in the dataset. This information in turn allows users to make inferences about the collection's contents and how they were generated and organized, and to learn about its characterization information—specifically, file format information and the corresponding preservation risks.¹⁷ Ultimately, this makes it possible to approach diverse data in the form of more manageable aggregated groups and facilitates long-term preservation decisions about what needs to be re-organized, labeled, re-named, or deleted if exact duplicates exist.

This phase of our collaboration began with the transfer to a single server of 1,370,000 files from the Metaponto collection, originally dispersed over several separate personal computers and external storage devices. The data, haphazardly organized depending on the computer or storage device of origin, were kept in their original order with reference to the original directories.

After the files were consolidated on a single server, DROID (the Digital Record Object IDentification tool developed by the UK National Archives)¹⁸ was run over the entire collection to identify file formats and checksums for the enormous number of digital objects it contains. Through a complementary script, the directory path to each file, the date of its last modification, and its file size

¹⁶ Weijia Xu, Maria Esteva, Suyog Dutt Jain, and Varun Jain, "Analysis of large digital collections with interactive visualization," in *Proceedings of the IEE Conference on Visual Analytics Science and Technology*, Providence Rhode Island, U.S.A. October 23 – 28 (2011), accessed September 14, 2012, doi:10.1109/VAST.2011.6102462.

¹⁷ Maria Esteva, Weijia Xu, Suyog Jain Dott, Jennifer Lee, and Wendy K. Martin, "Assessing the preservation condition of large and heterogeneous electronic records collections with visualization," *International Journal of Digital Curation*, 6, no. 1 (2011), doi:10.2218/ijdc.v6i1.171.

¹⁸ "DROID," accessed September 14, 2012, <http://droid.sourceforge.net/>.

were also obtained for all objects in the collection. These metadata could then be pre-processed and loaded into the visualization platform, allowing large amounts of information to be aggregated at more manageable levels. Different file formats, for example, are classified according to PRONOM's¹⁹ file format classification criteria: .jpg and .tif/.tiff, together with other known vector and raster file types, are classified as images, while .rtf, .doc and .docx files are classified as word-processor documents. In turn, checksums are processed to identify duplicates, and dates are aggregated per a number of years specified by the user. Different interactive functionalities make it possible to aggregate or select different metadata values, which can then be visualized in a way that allows the viewer to make inferences about the collection's contents and organization.

The pre-processed metadata were rendered visually as a treemap to show the collection's overall structure and the distribution of different file classes across the structure. The most obvious example of this is the use of the visualization tool to separate general administrative documents from research data (**Figure 3**). This was an important step towards identifying sections of the collection that are of highest priority for reorganizing. Using the aggregator function of the visualization (on the right in **Figure 3**), the user can find all the directories containing GIS data, which—since GIS files and administrative documents almost never occur together—allows in turn the implicit identification of directories that might be primarily administrative, and thus of low priority for reorganization. In this visualization, the user can see that GIS files are distributed in most directories, with the exception of the ones highlighted in white. By the same token, using the selector function (on the left in **Figure 3**) the user can visualize directories containing a very low proportion of image, database, and GIS files, which are the main file classes generated by staff whose primary function is research. That the directories in white in the first visualization appear again as directories with few image or database files (in black) in the second confirms the impression that those directories are not focused on research data. Conversely, the directories in which the largest proportion of files were pdf, email, word-processing and other text documents were deemed most likely to be administrative. The use of the visualization tool therefore allows the user to identify quickly and efficiently those parts of a collection that can safely be assigned a low priority for further intervention (in this case, administrative files as opposed to research data).

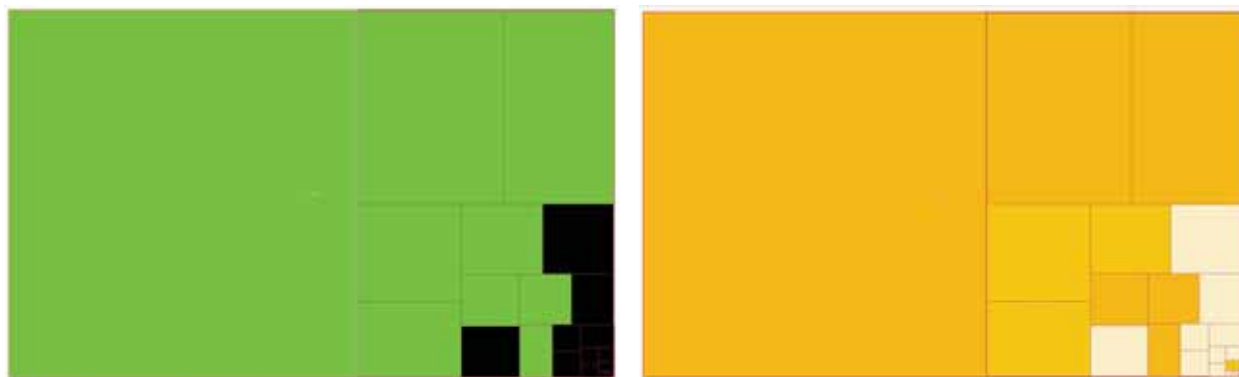


Figure 3. Location of administrative vs. research data using visualization. The image on the left shows ICA's GIS data identified via the aggregator function in the visualization. On the right is a view of directories containing a combination of GIS, image, and database data using the selector function.

¹⁹ "PRONOM," accessed September 14, 2012, <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>.

The visualization tool was deployed in this case to allow researchers involved in the Metaponto data triage to explore and make sense of a dataset generated over a long period of time by a large and revolving number of staff, specialists, and students. This situation is not uncommon for large academic archaeological projects that are carried out over long periods of time and with a high turnover in personnel. Visualization tools that provide an efficient and comprehensive way to understand what a disorganized collection is composed of, and how it might be reorganized, are therefore likely to find an eager user base among archaeologists. Using as our starting point the characteristics of ICA's data and the priorities identified in the initial triage of the Metaponto collection, we are currently developing new visual analysis functions that we hope will be applicable to other humanities collections as well. The new functionalities will allow us to re-establish control over the data, discard data that are redundant, and derive new information about the collection that can be integrated into subsequent documentation and preservation strategies.

4. Automated documentation, semantic web, and complex humanities collections: goals for the future

The Metaponto test case is itself a preliminary stage for a larger project with broader applications for other humanistic disciplines. It has become increasingly evident to us, as we have worked on this material, that most scholars are generating large quantities of digital information, much of which might turn out to be another scholar's research data in the future, and most have neither the time nor the skill-sets that will allow them to document their collections for preservation and reuse. In order to address this problem we are planning to re-engineer the tools that we used in our project to make them more flexible so they can be integrated to different standards and infrastructures. Semantic web principles will be an important part of this effort, since they offer the flexibility and potential for interoperability that are crucial in current efforts to ensure that digital data remain accessible and reusable.

The TACC-ICA team is therefore now concentrating its efforts on the development of plans for a project to create a desktop toolkit for the generation, extraction, and management of metadata for both structured and unstructured digital collections. This project will allow us to generalize our own experiences with archaeological datasets to the broader world of digital humanities collections. The basic functionality of the toolkit is a direct offshoot of our collaboration, on both a human and a technical level: simply put, it is meant to act as a translator between the disciplinary language of the collection creator and that of the digital archivist. It will allow the user to describe a collection intuitively, according to the intellectual framework and concepts with which he is familiar. At the same time, the software will encode those concepts, together with technical metadata extracted from the files, as structured metadata in formats such as XML that will be compatible with existing schemata and standards used by digital archives. Moreover, because the toolkit will allow researchers to manage their files and easily create standardized metadata both during and after the research process, it will ensure that data are more thoroughly and consistently documented from the beginning.

The toolkit will be built on the open-source ontology editor Protégé-OWL,²⁰ which will act as the back-end of a graphic user interface that will allow users to create organizational structures using the visual

²⁰ "The Protégé Ontology Editor and Knowledge Acquisition System," accessed September 14, 2012, <http://protege.stanford.edu/>.

metaphor of the “bucket”²¹ to replace the formal ontological concepts of classes and instances. To the user, the toolkit will present a seamless Graphic User Interface (GUI), in which he can extract metadata from his collections and drag and drop his files into user-defined categorical “buckets”. For example, an oral history institute could organize its collection by individual oral history cases or according to broader classes such as regions, localities, themes, etc. If a user wants to capture the process history of individual files, on the other hand, he could begin by creating buckets that reflect research stages, and within those, other buckets that represent units of observation. In the background the tool captures the bucket designations, and the relationships between buckets, in a standardized format within a formal ontology.

Behind the scenes, a mapping framework will allow a bucket labeled by a user as a unit of observation (e.g., stratigraphic context, oral history project, or work of art) or as a research stage (data gathering, analysis, interpretation) to be related to a standard metadata element in the ontology. In this way, the specialized concepts of ontology building are disguised for the user, who instead simply creates an intuitive organizational structure that is consistent with his own research workflow. To address the authenticity and integrity of the managed collection in compliance with archival and preservation requirements, the toolkit will create notations in the underlying ontology that record the presence, location and deletion of or change to the files in the collection as PREMIS OWL²² events when the user arranges and rearranges those files in the buckets. This document will also include information such as checksums, file format identification and dates of most recent modification, which will be extracted automatically from the files themselves. Technical metadata extracted from the files will be recorded in PREMIS OWL to provide information necessary for their long-term preservation. As end products, the ontologies built with the software will be exportable as XML/RDF and other metadata standards. In this way, at the end of the research project, both data and metadata can be integrated to digital repository infrastructures such as Fedora, which also uses OWL to establish relationships between properties of objects and other objects.²³

Such a toolkit will help to bridge the gap between data producers and repository managers by making it cheaper and easier for the producers to meet the curators half-way. Putting automated metadata creation tools in the hands of those who are responsible for digital datasets of any scale right now, without imposing rigid standardization or insisting on ingest into a specific repository, will make it simpler to apply repository-based preservation and dissemination solutions in the future. These tools will be especially important for digital collections in developing countries, where more permanent long-term preservation infrastructure does not yet exist. In the case of many digital humanities and cultural-heritage collections, the digital “memory of the world” is actually composed of the very human memories of individual researchers, many of whom have not yet had time to encode the knowledge in their heads as formal metadata for their digital collections. Without tools to facilitate the creation of such metadata, the world’s digital memory will diminish, little by little, with the inevitable loss of the human memories of each of those individuals. Where the digital record is the primary record of research, as it is increasingly in field archaeology, the resulting amnesia will be absolute and permanent. It is critical, then, that the digital preservation community address not only the preservation and dissemination of documents through digitization, but the preservation of meaning in the digital documents themselves.

²¹ Susan Cisco, “Big buckets for simplifying records retention schedules,” ARMA International’s Hot Topics. Trimming your Bucket List, *Supplement to the Information Management Journal* (2008): 3-6.

²² “Public workspace for PREMIS OWL ontology,” accessed September 14, 2012, <http://premisontologypublic.pbworks.com/w/page/45987067/FrontPage>.

²³ For example, the discussion of the description of relations using OWL in the Fedora ontologies wiki: “Ontologies - Fedora Repository Development,” accessed September 14, 2012, <https://wiki.duraspace.org/display/FCREPO/Ontologies>.

It FITS the Cultural Heritage!

Formats for Preservation: From Spatial Data to Cultural Resources

Giovanni Michetti¹ and Paola Manoni²

¹*The University of British Columbia;* ²*Vatican Library*

Abstract

Long-term preservation in digital environment is a complex process whose success depends on multiple factors: organizational, economical and technological issues must be carefully evaluated and balanced to design an effective preservation strategy. Formats are a relevant component of this strategy; therefore it is fundamental to analyse their properties. Conflicting factors need to be balanced in order to identify the best option for a specific environment or project, carefully weighing the different criteria adopted for assessment. The case-study of the Vatican Library and the digitization of its manuscripts collection is focused on a major challenge: the conversion of the digitized images to the Flexible Image Transport System (FITS), a non-proprietary format developed by NASA and long used for the preservation of geo-spatial data, as well as in astrophysics and nuclear medicine. The Vatican Library, in conjunction with the International Control Authority for FITS, is working towards the definition of a set of FITS header keywords describing cultural resources.

Authors

Giovanni Michetti is Assistant Professor at the University of British Columbia, where he teaches Archival Science. His research area is focused on contemporary archives, and his main research interests are records management, archival description and digital preservation. He has joined several international initiatives on digital preservation. He is deeply involved in standardization processes, as a President of UNI Sub-Committee “Records Management”, and Deputy of UNI Technical Committee “Documentation and Information” (UNI is the Italian Standards Organization). He is the Italian representative in a few ISO Working Groups on archives and records management. He authored several articles and essays.

Paola Manoni is librarian, metadata specialist and coordinator of IT Services at the Vatican Library. She is also a member of ISO Technical Committee 46 “Information and Documentation”, ISO liaison officer to IFLA, Deputy of UNI Technical Committee “Documentation and Information” (UNI is the Italian Standards Organization), President of UNI Sub-Committee “Technical Interoperability”, member of the EURIG (European RDS Interest Group) and an expert in IFLA Meetings of Experts on an International Cataloguing Code for the Statement of International Cataloguing Principles. Her research area is focused on metadata, digital libraries and frameworks of interaction systems.

1. Introduction

Scientific literature provides many different definitions of the term *format*, and possibly more different types of format classifications. Here we'll assume the PREMIS definition: a format is a *specific, pre-established structure for the organization of a digital file or bitstream*. As it regards the classifications, we'll refer to the well-known taxonomy of the Library of Congress, according to which formats may be evaluated on the basis of two different kinds of factors: 1) *Sustainability factors*, which “influence the

feasibility and cost of preserving content in the face of future changes to the technological environment,” and 2) *Quality and functionality factors* that vary by content category.¹

The sustainability factors are often used to assess the value of a format. The assessment matrices are all similar: the list of the relevant features on the left, and the assessment of these features in relation to a few formats on the right. We may even assign grades, so that at the end we’ll come up with a sum that will tell us what’s the best format. We may also improve this model, for example introducing weights. Anyway, we are missing something, if we adopt this approach. What needs to be clear is that actually these factors are not independent parameters. They need to be analysed and weighed in relation to the overall preservation framework. In other words, we want to remark that the original statement provided by LOC states: “These [sustainability] factors are *significant* whatever strategy is adopted,” which means they are relevant, and must be analysed and evaluated. Our impression is that this statement has been widely understood as: “These factors are *required*, whatever strategy is adopted.” Let’s step back and look at the larger framework.

Digital preservation is the active management of digital objects over time to ensure ongoing access. This quite generic definition is fine as long as we understand that physical (bit) preservation is not at all the issue: preservation implies migration or anyhow an alteration of the original object. Intellectual preservation is the real problem because it entails preserving the intellectual content in the face of future technological and knowledge changes. So the problem is to understand what is the *meaning* of an object: what are the features that provide meaning to an object? What is the context that gives meaning to an object? To what extent do we need to consider the context of an object? In brief, it is not just a matter of data longevity; it is rather a matter of knowledge/context longevity. That’s why digital preservation—rather, *long-term* digital preservation—is a complex system of activities, methodologies, tools, and other entities, in which we may distinguish:

- *What* we are preserving
- *Why* we are preserving
- *Who* is involved in the preservation process
- *Where* we are performing preservation
- *When* we put in place preservation activities
- *And how*.

The *how* question leads to organizational issues, hence to procedural, economical, and technical issues. Here we find the *formats*. So the format is part of the problem, but it has to be seen as a technical issue related to all these different profiles of digital preservation.

Let’s take for example *stability*: this feature is usually required for any sustainable format. *Stability* means that the format shouldn’t be “subject to constant or major changes over time.” Actually there are very few formats that are really stable according to this definition, since the majority of formats evolve constantly. Anyway, our concern here is: stability, a relevant feature when we talk about *long-term* digital preservation? We guess that evolution will be managed easily within any digital repository in charge of *long-term* digital preservation, according to pre-defined procedures. In other words, if we deal with *long-term* digital preservation, we may take for granted that in such a long range of time there will be a few changes of format, therefore, stability doesn’t seem a relevant issue.

¹ See <http://www.digitalpreservation.gov/formats/>.

So the question is—features like stability are desirable indeed, but to what extent? To what extent are the sustainability factors identified by the LOC required? The overall framework presented before may help in such evaluation. Let's see some examples.

2. Adoption

“Degree to which the format is already used by the primary creators, disseminators, or users of information resources.”

It seems a reasonable requirement: the more a format is used, the more chances of preservation we have. But let's go back to late Nineties, and consider XML: there has been a time when XML was not a widespread adopted standard. It's even worse if we consider SGML. And what about RDF today? We can't say it's a well adopted format. Be it SGML, XML or RDF—may we state that preserving objects in one of those formats was/is a wrong choice? We guess not. Wide adoption is not a positive value per se: it must be analysed in context. Here, the *why* is relevant: do we want to preserve the material and at the same time implement up-to-date tools to query and locate the material? RDF may be a choice. But it may be an expensive choice, so the *how*—the economical factor in particular—is relevant as well. The *who* may be relevant too: staying on RDF, we know that implementing RDF may allow users to perform more sophisticated and meaningful queries, so the problem is, what kind of searches are we planning to perform on the objects? *Who* is the user we are targeting? It may require specialized skills to implement and maintain RDF architecture: *who* is in charge of this function? Do we have this skill in house, or do we have to search elsewhere?

3. Transparency

“Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor.”

Transparency is usually inversely proportional to efficiency:² for example, XML files and raster images have a high degree of transparency but they are not very efficient. The less a format is transparent, the more it depends upon compliant software, tools and algorithms to read/edit, and the more it requires sophistication to build tools. Hence, the transparency requirement is strongly related to organizational and economical issues (the *how*): again, it is not a positive value per se. If an object has to be preserved for the long-term, to what extent do we care of ease of building tools? Aren't we more interested into saving space, especially if we are managing big amount of data, if not big data? Also, the *who* may have some relevance: in specific domains (e.g., in astronomy) couldn't we assume that stakeholders have a deep and sophisticated knowledge of the preserved objects along with their features, no matter of transparency? So, transparency has to be considered in relation to the overall framework.

² The extent to which space is well used.

4. Self-Documentation

“Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata.”

Self-documenting formats seem the best option in any case, because we can associate useful information to an object embedding metadata into it. But everything comes to a cost: embedding metadata into a file means increasing its size. However, not all metadata are specific to individual files: non-specific, shared metadata may be put *out* of the file, as a metadata package associated to a bunch of files, and redundancy would be reduced. It could be argued that embedding is admittedly a safer option than associating but it has to be remarked that 1) as we said before, preserving an object means preserving the context, so we cannot avoid associations, we'll have to deal anyway with a network of associated objects providing meaning to the preserved object, and 2) embedding metadata leads to a file size increase, which in turn may slow down the access to the file. Here we are again: this sustainability factor needs to be assessed in the context of a preservation project. The problem is not to select the “most self-documenting” format, it's rather to select a format whose self-documentation features fits our preservation strategies, procedures, policies.

5. External Dependencies

“Degree to which a particular format depends on particular hardware, operating system, or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.”

Like self-documentation, in principle we would prefer a so-called *self-contained* format, one in which we can find all the dependencies, the knowledge needed to interpret and understand an object. Such an object would have more chances to *survive* if it *gets lost* in the real world, because in the future we'll still be able to understand them. But in the real world we develop and implement long-term digital preservation *projects*, we deal with a *systematic* approach to digital preservation. This means adopting a framework³ in which any dependencies are made explicit. The typical example is PDF, whose faithful rendering requires that fonts be embedded: is it better to embed the fonts in the file, or preserve the fonts as a separate file and associate it to individual files? The solution—and the selection of the most adequate format—have to do with the *how* and the organizational issues again. The *who* is also relevant: we would suggest a self-containing format to a generic user, because we don't expect them to perform a skilled and thorough action aimed at long-term preservation. But when dealing with a specialized, trustable custodian, we believe that proper strategies and procedures are put in place, and we may think about avoiding a self-containing format.

Similar considerations hold for other sustainability factors, such as disclosure, complexity, interoperability. A further level of complexity may be added, if we introduce the authenticity issue.

The rationale behind these considerations is that all these features should not be interpreted just as *sustainability* factors. Rather, they are factors that *need to be evaluated when aiming at sustainability*. And this evaluation must be done taking into account the overall framework as well as the fundamental

³ For example, OAIS or COP (Chain of Preservation Model, developed by InterPARES).

requirement for any preservation action, i.e., the preservation of authenticity. The assessment of these factors is not absolute; rather it strongly depends on the context.

Hence, there is no one-size-fits-all format for digital preservation: conflicting factors need to be balanced in order to identify the best option for a specific environment or project, carefully weighing the different criteria adopted for assessment. From a careful evaluation, we can find the format that best fits our needs. From a careful evaluation, FITS has been found as *the* format the Vatican Library needs for its digitization projects.

Manuscripts and antique books of the Vatican Library's collections are being digitized to preserve them for future generations adopting a file format developed in the context of space missions and storing satellite images of the sky during the end of 70s of the last century: the Flexible Image Transport System (FITS), an open format, fully documented, without royalties or copyright, based on a series of specification publicly available and managed by a non-profit scientific authority.

The challenge of the Vatican Library preservation project is the conversion of the digitized images to the Flexible Image Transport System (FITS) with the aim to preserve images by using the long-lived technology evolved out over decades in the astronomical field and with an experimental approach to adjusting this format for libraries' purposes with the belief that:

An archival format must be utterly portable and self-describing, on the assumption that, apart from the transcription device, neither the software nor the hardware that wrote the data will be available when the data are read.⁴

But let me give you a preliminary insight on the history of this format. In the early 70s the number of space missions in the world was growing, but almost every mission had its own format for storing data and images collected. In order to avoid this inconvenience, the scientific community studied at length the problem and began the implementation of a common *format*.

FITS is the standard computer data format widely used by astronomers to transport, analyse, and archive scientific data files when it became necessary to settle a procedure for transferring astronomical data from one installation to another.⁵ The first version of FITS, in 1979, consisted of a binary array, usually multidimensional, preceded by an ASCII text header with information describing the organization and contents of the array. The FITS concept was later expanded to accommodate more complex data formats.

Then in 1982 the International Astronomical Union (IAU) formally endorsed the format. Provisions for data structures other than simple arrays or groups were made later. These structures appear in extensions, each consisting of an ASCII header followed by the data whose organization it describes. The IAU General Assembly approved a set of general rules governing such extensions and ASCII table, in 1988. In the same year, IAU FITS Working Group (IAUFWG),⁶ the international control authority for the

⁴ National Research Council (U.S.), Steering Committee for the Study on the Long-Term Retention of Selected Scientific and Technical Records of the Federal Government, *Preserving Scientific Data on our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources* (Washington, D.C.: National Academy Press, 1995), 60.

⁵ See FITS Standard Document. http://fits.gsfc.nasa.gov/fits_standard.html.

⁶ Currently there are 23 members of the IAU-FWG, chosen to represent the interests of both ground- and space-based astronomy at all frequencies that astronomers observe, and to represent various regions of the World, major astronomical data centers, major astronomical software packages, and the historical traditions of FITS. The current

FITS, was formed under IAU Commission 5 (Astronomical Data) with the mission to maintain the existing FITS standards and to review and maintain future extensions to FITS, recommended practices for FITS, implementations, and the thesaurus of valid FITS keywords. In 1989, the IAUFWG approved a formal agreement for the representation of floating point numbers. In 1994, the IAUFWG endorsed two additional extensions, the image extension and the binary table extension.

The NASA/Science Office of Standards and Technology maintains the standard and its updated⁷ versions. In July 2008, the IAU FITS Working Group officially approved the new 3.0 version of the FITS Standard document with the guarantee the any structure that is a valid FITS structure shall remain a valid FITS structure at all future times.

A FITS file is made of 2880-byte records called *FITS* blocks divided between a header and a data area. The major feature of the FITS format is that image metadata is stored in a human-readable ASCII header, so that an interested user can examine the headers to investigate a file of unknown provenance. Each FITS file consists of one or more headers containing ASCII card images (80 character fixed-length strings) that carry keyword/value pairs, interleaved between data blocks. The keyword/value pairs provide information such as size, origin, coordinates, binary data format, free-form comments, history of the data, and anything else the creator desires: while many keywords are reserved for FITS use, the standard allows arbitrary use of the rest of the name-space.

Keywords may appear in any order except where specifically stated otherwise in this standard. It is recommended that the order of the keywords in FITS files be preserved during data processing operations because the designers of the FITS file may have used conventions that attach particular significance to the order of certain keywords (e.g., by grouping sequences of COMMENT keywords at specific locations in the header, or appending HISTORY keywords in chronological order of the data processing steps).

In comparison with other file formats, for example the TIFF format, the FITS file has not limitation for the reading of files of any size and every kind of numeric or textual data can be saved: integer or real numbers, 32 or 64 bit, matrices of any dimensions where images can be treated like bidimensional matrices where colors and encoding are the values of the matrix. Internal addressing data has not critical points, and the only limit is the size of files allowed by the operating system.

On the other hand image processing programs such as GIMP, Photoshop, XnView and IrfanView can generally read simple FITS images, but frequently cannot interpret more complex tables while almost all graphics programs or viewing photos can read and write TIFF files in each operating environment and special libraries allows the reading of these files from the most common programming languages.

The choice of FITS file as long-term preservation format implied the implementation of a conversion tool because the FITS file, in the Vatican digitization project, is not set as a publishing format *but only as storage file format*.

The Vatican Library, in collaboration with experts of the INAF (the Italian National Institute for Astrophysics), is working to an open source tool, able to perform lossless conversion of data with particular regards to the TIFF / FITS and vice versa conversion. In this way the Vatican Library is going to store manuscripts digital images in FITS file in its WORM data storage device, ensuring the availability of the other formats for the needs of scholars and end-users.

chairman of the IAU-FWG is William Pence (NASA/GSFC) and the vice-chairman is Lucio Chiappetti (IASF Milano, IT).

⁷ See The FITS Support Office, <http://fits.gsfc.nasa.gov/>.

This conversion is taking into account all the characteristics of technical metadata (for example in TIFF format) in order to identify matches in the keywords of the header portion of the FITS file.

The FITS community has developed many conventions for using certain keywords or FITS file structures.

The Registry of FITS Conventions provides a central and authoritative repository for documenting conventions that have been developed by the FITS user community for storing and transmitting various types of information in FITS format data files. A FITS convention is defined as a set of related FITS header keywords, and optionally, other data structures within FITS tables, FITS images, or other types of conforming FITS extensions that are to be used for a specific purpose. The IAU FITS Working Group is responsible for this Registry and for the rules and procedures for entering new conventions into it.

A process for the establishment of an ‘ad hoc’ convention based on the need for the handling of digital images of library holdings is under consideration. For instance, the assignment of new keywords for elements not already stated for astronomical images.

For example, FITS, unlike “picture” formats like JPEG and TIFF, has no native convention for storing colour information and it will be necessary to define new FITS convention for it. The overall idea of a new convention for library holdings data includes both the requirements of new keywords and the embedded linkage to the preservation metadata: the PREMIS file related to image files.

The so called “foreign extension” will encapsulate the PREMIS reference while the XML file will automatically include the contents of FITS keyword mapped in each related PREMIS data elements corresponding to the ‘Object’ entity.

In fact extensions to the basic FITS format can be defined as specified in the FITS Standard document. Following the model, the only restriction that will have to be placed on the freedom to create new extensions is that there should be only one approved extension format for each type of data organization. New extension types have to be created whenever the organization of the information is such that it cannot be handled by one of the existing extension types. It will be the function of each user who creates a new extension type to check with the standards committee to see if an extension already exists for that type of data organization and to propose one if it is really a new extension type.

‘Foreign extension’ type is used to put a FITS wrapper about an arbitrary file, allowing a file or tree of files to be wrapped up in FITS and later restored to disk. A full description of this extension type is given in the FITS Registry of conventions.

The essence of the idea to use FITS is as a very stable, well defined, easy to reverse engineer format for very deep storage over the decades while the concept of ‘foreign extension’, on the other hand, is a way to insert a third-party standard into a FITS stream. In the case of the PREMIS as a third-party standard is able to ensure the same stability as FITS, so that permanence won’t be sacrificed and the assertion: “any structure that is a valid FITS structure shall remain a valid FITS structure at all future times” will be still valid.

In conclusion we would like to point out that the Vatican Library, in making the choice of using FITS, has encountered willing collaboration on the part of researchers in the field of astrophysics, and at the same time it is now enjoying the valuable partnership of the international institutions which handle the FITS standard, with the goal of making it fully adequate for the preservation of library holdings. Even though the starting point of these areas of research, astrophysics and humanities, is very different, for this once it becomes so close to each other that they can share a digital conservation format for their respective images.

File Viewers

Examining On-the-Fly File Format Conversion

Lois Enns¹ and Gurb Badesha²

¹Records Manager, City of Surrey; ²Functional Application Specialist, City of Surrey

Abstract

File viewers are utility applications that identify file formats and render source files in human-readable form using on-the-fly file format conversion and without triggering a native application. While many archives follow a file format conversion strategy for long-term digital preservation, other organizations may experience significant barriers to this preservation strategy in terms of resources, technology, risks, and drivers. Working within an InterPARES 3 general study, co-investigators at the City of Surrey tested six file viewer products to answer four research questions: how do file viewers work; what software is available for use; how accurately do file viewers render files; and what role might file viewers play in digital preservation. Based on test results, opportunities were identified for using file viewers as a component of a digital preservation strategy to reduce resource requirements, extend backwards compatibility, and improve electronic appraisal procedures.

Authors

Lois Evans has a Master of Information Studies from the University of Toronto, and has worked in local government for nine years as a district archivist and records manager, now with the City of Surrey. As a co-investigator on the UBC InterPARES 3 project, Lois developed an end-to-end procedure for appraising files on shared drives and migrating records to an electronic records management system, and published the *Shared Drive Migration Toolkit*. Lois has written on e-government and e-records for a number of publications.

Gurb Badesha has a Bachelor in Interactive Arts and Technology from Simon Fraser University, and has worked in local government at the City of Surrey for three years as a records coordinator and functional application specialist on the electronic records management system. Gurb participated in the shared drive migration project, and wrote the *Utility Applications Guide* which accompanies the *Shared Drive Migration Toolkit*.

1. Background

During the UBC InterPARES 3 case study on developing a production-oriented procedure for appraising and migrating files from shared drives to an electronic content management (ECM) system (Rogers *et al.*, 2010), the InterPARES co-investigators at the City of Surrey identified and adopted a number of utility applications to expedite our work. These utility applications included: a disk space manager, used to collect drive statistics, analyse file formats, create historical profiles, and facilitate metadata discovery; a file manager, used to apply unique identifiers and rename records; a duplication finder, used to identify and remove byte-by-byte duplicates; a format identifier, used to identify and resolve missing file extensions; and an empty folder identifier, used to count and remove empty folders. These activities are described in the *Shared Drive Migration Toolkit* (Enns and Badesha 2011). Although over 285 file formats were identified during the course of the project, only 47 file formats were confirmed as records suitable for migration, and only two of these file formats were found to be obsolete. These two file

formats (.ptn and .dwt) represented only 18 files out of 98,197 selected for migration. The remaining 45 file formats could be opened using available native applications.

The Surrey case study did not address digital preservation. Although many of the migrated records were scheduled for permanent retention, conversion to preservation formats was determined to be out of scope for the project, due to a number of barriers. Resource constraints included lack of disk storage space, staff capacity, staff time, and lack of documented standard operating procedures. Technical difficulties included a lack of capacity to manage bulk conversions and related metadata in either the source or target environments. Additionally, none of the conversion drivers identified in *ANSI/ARMA 16-2007 The Digital Records Conversion Process* (American National Standards Institute 2007)—such as retention requirements, operational factors, or regulatory or legal factors (p. 4)—appeared to fit the situation. Finally, the risk of “degradation or loss of the accuracy, completeness, authenticity, and integrity of the records” (p. 1) for format conversion appeared high, considering in the migration work already underway. For these reasons, both the records management and the information technology teams were reluctant to commit to a conversion strategy at this time.

As well, the records team were aware that the file formats in the Surrey environment appraised for migration were not necessarily subject to immediate technical obsolescence, since only two formats and 18 files were obsolete. Given that the vast majority of the files were not obsolete and did not appear to be under threat of obsolescence, the records team wondered whether the question of file conversion might be postponed indefinitely. Around the same time, the records team tested an ECM-integrated file viewer module that allowed users to open and annotate specialty drawing files (i.e., .dwg) without using the native application (i.e., AutoCAD). Although subsequent testing revealed that the module was not well-integrated to the ECM system (and it was not adopted), the idea that a file viewer might somehow extend the life of a file format was appealing.

As a secondary consideration, the records team found that during file appraisal activities, opening files to validate contents was a time-consuming activity. Only a few applications could be effectively managed on a computer task bar, and time was spent waiting for applications to open and files to load, and in flipping between native and utility applications. A file viewer supported multiple formats from a single point was worth pursuing.

In May 2011, the InterPARES 3 Team Canada members approved a general study on file viewers. Four areas of interest were identified: how do file viewers work; what software is available for use; how accurately do file viewers render files; and what role might file viewers play in digital preservation. Over the course of the next year, these questions were examined by the two Surrey co-investigators, with participation by two graduate research assistants, and input from members of Team Canada at bi-annual workshops. Study activities included: a literature review; correspondence with file viewer developers; selection of file viewer products for testing; development of basic product comprehension and creation of a test environment; identification of file formats, properties, characteristics, and files for testing; testing of products and collection of data; and examination of results.

2. Literature Review

A number of articles mentioning file viewers are found in software and computer engineering journals, primarily with respect to the role of file viewers in software design. For example, an article on a product called GroupKit mentions a file viewer in the context of enabling users’ views of text documents in a conferencing environment (Roseman and Greenberg 1995, p. 6). Other articles mention file viewers in the

context of software programming, along with other types of viewers: a directory viewer, an error viewer, an execution viewer, a software landscape viewer, and an interface viewer (Manoridis *et al.* 1993 pp. 16, 18) and a project viewer and a graph viewer (Anderson and Teitelbaum 2001, p. 3). Evidently, file viewers are one of a number of viewers used to interpret machine language into human-readable form.

Adjacent to this work are articles on file format identification, a component of file viewing. There are at least three computer-based methods for determining file formats: extension-based detection; magic-numbers-based detection; and content-based detection (Amirani *et al.* 2008). Essentially, the extension-based approach uses file names and mime types; the magic number approach uses the “secret” numbers hidden in file headers; and the content-based approach references “fileprints” through different types of frequency analysis (McDaniel and Heydari 2002 and Amirani *et al.*). Scattered through these technical articles are suggestions as to why file format identification work is important, including: detection of changes made by a malicious user; dealing with proprietary file types; obsolescence (Dhanalakshmi and Chellappan 2009); and the need “to preserve data beyond the life of a particular piece of software” (McHenry *et al.* 2009).

Within the format-identification articles, “Towards a Universal, Quantifiable, and Scalable File Format Converter” (McHenry *et al.*) is of particular interest. Here, the authors express concern that since “not every format supports the same data content” (p. 140), data is dropped when a file is converted from one format to another. In order to minimize the data lost during conversions, they propose a “polyglot,” or “a framework for measuring the quality of individual conversions and allowing for the use of this information in choosing optimal conversion paths” (p. 146). They note that, “Aside from the ability to convert between many formats another useful application of such a potentially ‘universal’ converter is in the form of a ‘universal viewer.’ Given the ability to view one format in each domain, one could potentially view them all with such a converter by converting every file to this target format...” (p. 146). With many archival and records institutions following conversion and/or pathway strategies for long-term digital preservation, a universal file viewer that converts source formats to destination formats “on the fly” presents intriguing new possibilities.

Focusing on file formats, a number of articles and project reports in the library and archives realm examine the significant properties of file formats or “the characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects” (Wilson 2007a, p. 15). Many digital preservation projects (e.g., Investigating Significant Properties of Electronic Content over Time [InSPECT], Creative Archiving at Michigan and Leeds Emulating the Old On the New [CAMiLEON], Consortium of Research Libraries Exemplars in Digital Archives [CEDARS], Preservation and Long-term Access through Networked Services [PLANETS]) and national archives (e.g., National Archives of Australia, National Archives and Records Administration [US], The National Archives [UK]) have published papers or web articles on significant properties, also called “significant characteristics” or “essential characteristics”. Significant properties provide a means of measuring whether a preservation strategy such as migration or emulation is successful, by comparing how well a target file retains the properties found in the source file. The “Significant Properties Report” (Wilson 2007b) provides a useful overview, beginning with a reference to “Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information” which notes, “We want to be able to guarantee that for a given object the reformatted version is equivalent to the original version with regard to some specific set of object characteristics” (Lynch as quoted in Wilson 2007b, p. 5).

An important shift in the significant properties discussion came with the general acceptance that digital objects “do not need to remain in a state that is unchanged from their original state in order for

them to be considered authentic” (Wilson, 2007b, p. 4). Instead, “A record is considered essentially complete and uncorrupted if the message meant to communicate in order to achieve its purpose is unaltered” (as quoted in Wilson 2007b, p. 4). However, there is an ensuing problem as what is considered “essential” varies from audience to audience. For example, when looking at medieval manuscripts, an audience interested in text analysis would consider the text of a document to be essential, while an audience interested in literary metaphor would insist that the illustrative and design components as important as the text. Despite a “pressing need” to “develop a methodology, and begin identifying quantifiable sets of significant properties for specific classes of digital object[s]” (Wilson 2007b, p. 7), there is no definitive set of significant properties available. Although some studies provide examples of significant properties for audio, email, raster images, and structured text (Grace 2009), the *InSPECT Framework Report* reflects a general move towards developing a methodology or framework whereby “an evaluator operating in a curatorial institution can determine the properties that they consider to be essential based on their interpretation of acceptable loss” (Knight 2009, p. 9). To this end, institutions such as the Library of Congress and the Florida Digital Archives have identified and posted the significant properties of interest referenced by their institutions on their websites.

3. Methods

Following the literature review, the co-investigators selected file viewer products for testing. Two categories of file-viewer software emerged: low-cost file viewers intended as stand-alone products; and more costly file viewers intended for integration with other software. This study focused on low-cost, stand-alone products costing less than \$100 per license. Ease-of-use and the number of format categories covered by the product were two other important criteria. A number of Google searches were completed (e.g., “file viewers,” “universal viewers,” “best file viewers”) and a preliminary list of products was identified.

Next, the products were qualified using Download.com, a site featuring software reviews, technology news and software downloads, and SourceForge.net, a site for open-source software development. Once the products were short-listed, each product website was reviewed to identify the best fit for the project, and the final product selection was made. Although open-source file viewers were identified, only one open-source file viewer supported two of the six format categories, and an attempt to download this product was unsuccessful due to programming requirements. In the end, the products selected for testing included: Accessory Software File Viewer (\$23.00); FileStream Turbo Browser (\$69.00); GetData Explorer View (\$29.95); Irfan View (\$10.00 donation); Quick View Plus (\$49.00); and UV ViewSoft (\$25.00). Once the products were selected, the co-investigators contacted the developers using email and web forums to ask questions about how file viewers work. In every case, the developers were advised that the co-investigators were seeking information for a research paper on file viewers. Most of the developers replied, and sufficient information was provided to create a general understanding of how file viewers work.

A test environment was set up to host the six file viewer products. The environment included two workstations: a Windows-platform workstation connected to Surrey’s networked computing environment; and a Windows-platform personal laptop owned by one of the co-investigators and not connected to the network. All of the test files were maintained on the Surrey workstation, and all of the file viewers were downloaded to the personal laptop. The test files were transferred from the workstation to the personal laptop using a USB drive. Once the six file viewers were loaded to the laptop, the co-investigators spent some time orientating to the products, and eventually ran a complete set of test files to confirm their

understanding of the products and testing routine. The test run included seven file viewers (including one trial version later not adopted), 14 file formats, and nine files for each format, with three files selected from three time blocks (1994-1999; 2000-2005; and 2006-2011) to test whether file viewers are to any degree backwards compatible.

With the test environment and file viewers in place, the co-investigators looked for ways to measure how well the file viewers rendered files, referencing the significant properties listed on the InSPECT, Florida Digital Archives, and the Library of Congress websites for each format category. Here, the co-investigators took a somewhat different approach, separating significant properties into two somewhat arbitrary groups: **properties**, which could be determined without opening a file; and **characteristics**, which could only be determined by opening a file. For the purpose of this study, properties represent metadata that can be reviewed using a disk space manager, while characteristics represent metadata that cannot be viewed using the disk space manager as well as content. In a best-case scenario, the two groups would be separated into **metadata properties** and **content characteristics**, where properties include all metadata and characteristics reflect content alone.

For all format categories, three properties were consistently identified: Title, Creator, and Date Created. Additional properties were identified by file format category: Word Count (for text); Resolution, Bit Depth, Width, and Height (for images); and Length, Width, Height, Pixel Aspect Ratio, and Frame Rate (for moving images). These properties could be assessed using the Windows operating system and/or a file manager utility application, and the native application. Property data was collected and reviewed (see Table 3) but did not play a part in determining how well file viewers render files.

Characteristics that could be assessed using file viewers included: Header and Footer, Font Size and Colour, Images/Diagrams, Bullets and Numbering, Print, *Hyperlinks*, *Page Count*, and *Text Search* (for text); *Font Size and Colour*, Cells, *Formulas*, *Macros and Links*, *Frames/Page Breaks* (for data); Font Size and Colour, Sender, Receiver, Name, Date Sent, Date Received, Subject, Attachments, Body, Signature (for email); Division, Paragraph, Image, Link, Frame (for web); Font Size and Colour; Colour, Scalability, Sharpness, *Page Number* (for drawings); Colour, Completeness (for images); and Colour, Sound, and Back and Forward Navigation (for moving images). Mandatory characteristics (on which the later pass/fail assessments were made) are displayed in regular font, while *optional characteristics* are displayed in *italic* font. Some characteristics, such as slide presentation and animation (.ppt) or formulas, macros, and links (.xls) were not represented by any of the file viewers. The lack of conversion of these characteristics is also common to .pdf format conversion. These characteristics were treated as non-mandatory.

Of the 45 file formats migrated to the Surrey ECM system, only 12 file formats represented at least 500 files and up to 18 years worth of instances. These file formats became the focus of file viewer testing and included: .doc, .pdf, .ppt, .xls, .msg, .htm, .dwg, .vsd, .jpg, .tif, .mov, and .avi. While the selection of formats chosen by another organization might differ, the co-investigators felt that these formats were quite common, and represented formats they would need a file viewer to render if it was to be used in any appreciable way for production purposes. Once the formats were selected, significant care was taken to ensure that files chosen for testing presented the properties and characteristics of interest, and files were chosen from each of three time blocks (except for .avi, where only files from 2006-2011 were found). The testing was done twice, using two different sets of nine files for each of the 12 formats.

During preliminary testing, a discovery was made that four out of the six viewers could not render Microsoft files in the “x” file formats (i.e., .docx, .pptx, .xlsx), designed to meet the Office Open XML standard. Additionally, the file viewers could open .htm files but rendered the files as text representations with style tags, without graphic representation. The reason for the “xml” gap in the file viewers is not

known. Perhaps the developers of these products do not consider .xml file formats problematic, assuming that these files will be viewed using a web browser or editor. Or perhaps the .xml file formats are too new, and the developers have not had time to bundle in an appropriate viewer. At any rate, these file formats were removed from the test sample.

In addition, two test files could not be opened in the native application and were considered corrupted. These files were removed and replaced.

Once the files were selected and placed on the workstation and the laptop, the six file viewers were tested. Each file was opened on the workstation using the native application, and then on the laptop using the file viewer. Using a file format instance chart (see Table 4, 5), each characteristic presented on the laptop was compared to the workstation, and given a pass or fail. In total, 72 file format instance charts were completed.

Using the file format instance charts, a determination was made as to whether or not the file viewer successfully rendered the file format for the time block. A pass meant that all mandatory characteristics were successfully rendered (see Tables 6, 7, 8). The formats were then grouped, and the file viewer was given a pass or fail for the format category (see Table 9).

4. Results

The results are presented with reference to the four research questions posed for the IP3 General Study.

4.1 How Do File Viewers Work?

In general, file viewers work by identifying file formats through header information, magic numbers, or content, and then rendering the content in human-readable form. Some file formats are rendered “as is” from the source file, while others are converted on-the-fly from the source format to a target format that can be rendered. In order to extend their file format rendering capabilities, file viewers often consist of a number of viewers bundled together. For example, one respondent noted their product used viewers from Internet Explorer (for text, html, and Microsoft Object Linking and Embedding or OLE files); Leadtools (for image files); and Delphi (for data files with open database compliancy or ODBC), while another product leveraged the Microsoft Internet Explorer engine (for html files); a doc-rtf converter (for text files); and Delphi (for data files). A third respondent referred to “third-party libraries,” and a fourth noted the use of “outside-in” libraries which convert “foreign” formats to a generic format that leverages a standard viewer. As noted by one respondent, file viewers are “actually rendering a much smaller number of standard formats” than the 100 to 300 file formats commonly listed in their product information. The file viewer bundling approach was demonstrated during testing, when all six of the file viewers tested launched Adobe Reader to render .pdf files.

In some cases, the file viewer product is intended for specific format categories—for example, IrfanView is intended for use with image and audio/video file formats only, while FileStream Turbo Browser is intended for wider use and extends to six format categories. During the product selection phase, the types of format categories targeted by the products were captured (see Table 1). Based on this product information, the co-investigators expected that the FileStream Turbo Browser and Quick View Plus viewers would perform the best during testing.

Table 1. File Viewer Rendering Capabilities by File Format Type (based on product information).

Products	Text	Data	Email	Drawings	Images	Moving Images
Accessory Software File Viewer	Yes	Yes	No	No	Yes	Yes
FileStream Turbo Browser	Yes	Yes	Yes	Yes	Yes	Yes
GetData Explorer View	Yes	No	Yes	Yes	Yes	Yes
IrfanView	No	No	No	No	Yes	Yes
Quick View Plus	Yes	Yes	Yes	Yes	Yes	No
UV ViewSoft	Yes	No	No	No	Yes	Yes

The more costly file viewers intended for integration with other software often offered additional features in concert with file rendering: format conversion; editing; annotation, redaction, and integration. These features were less common in the lower cost file viewers investigated in this study (see Table 2).

Table 2. File Viewer Additional Features (based on product information).

Products	Format Conversion	Edit	Annotation	Redaction	Product Integration
Accessory Software File Viewer	No	No	No	No	No
FileStream Turbo Browser	Yes	Yes	No	No	No
GetData Explorer View	No	No	No	No	No
IrfanView	No	No	No	No	Yes
Quick View Plus	No	No	No	No	No
UV ViewSoft	No	No	No	No	Yes

4.2 What Software Is Available For Use?

As mentioned, available software falls into two categories: low-cost file viewers, intended as stand-alone products; and more costly file viewers, intended for integration with other software. The focus of this study was low-cost file viewers, and there are dozens products available, beyond the six products selected for this study.

4.3 How Accurately Do File Viewers Render Files?

As mentioned, a number of metadata properties were reviewed using a disk space manager and were not considered as part of the file viewer testing. These properties were collected in tables, reviewed, and set aside (see Table 3).

Table 3. File Format Properties (.doc).

DOC	Title	Creator	Date Created	Word Count
1994 to 1999				
File 1	Tracer Introduction and Configuration.doc	Administrators	1996-08-06 9:00	206
File 2	Instructions to Upgrading Firewall.doc	Administrators	1996-10-03 8:52	697
File 3	DCT CSDC Documentation Amanda 3.doc	SURREY\LSA	1996-08-26 10:49	5472
2000 to 2005				
File 1	DCT Audit Report Procure Audit Report.doc	SURREY\NAJ	2000-01-13 16:21	4293
File 2	Steps for Renaming Production databases.doc	SURREY\BL8	2002-07-09 14:12	2710
File 3	DCT Old Pre 7 4 Documents Cognos 1.doc	SURREY\IAM	2000-01-17 07:10	363
2006 to 2011				
File 1	DCT IP3 Creator Preserver Responsibilities V 03 0.doc	SURREY\LE2	2009-05-22 13:03	3188
File 2	SOW Storage Solution Facilities Plans 2008 08 25 v01 0.doc	SURREY\LE2	2008-08-26 07:27	935
File 3	DCT Master List 2011.doc	SURREY\EAG	2010-12-02 11:00	3051

Next, file format characteristics of the files were compared using the native application rendering of the source file on the workstation, and the file viewer rendering of the file on the laptop. These results were recorded in 72 file viewer instance charts (i.e., six file viewers x 12 file formats). Nine files were tested for each format, with three files from each time period (i.e., 648 files). The characteristics were charted, with mandatory characteristics in regular font and *optional characteristics* in italics (see Table 4).

Table 4. File Viewer Instance Chart Showing Pass Results (.doc).

ACCESSORY SOFTWARE FILE VIEWER									
DOC	Header/ Footer	Font	Images/ Diagrams	Bullets	Hyperlink	Page Count	Text Search	Print	
1994 to 1999									
File 1	PASS	PASS	PASS	PASS	N/A	PASS	PASS	PASS	
File 2	PASS	PASS	N/A	PASS	PASS	PASS	PASS	PASS	
File 3	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	
2000 to 2005									
File 1	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	
File 2	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	
File 3	PASS	PASS	PASS	PASS	N/A	PASS	PASS	PASS	
2006 to 2011									
File 1	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	
File 2	PASS	PASS	PASS	PASS	N/A	PASS	PASS	PASS	
File 3	PASS	PASS	N/A	PASS	PASS	PASS	PASS	PASS	

For each characteristic, the co-investors compared the native application rendering of a file with the file viewer rendering of the same file, and marked the file viewer characteristic with a pass or fail. Based on the mandatory characteristics, the file viewer passed (see Table 4) or failed (see Table 5). Although a file viewer was given a fail if just one mandatory characteristic failed, in most cases, the results of the test were fairly obvious, with a number of fails noted (see Table 5). If the characteristic was not present, it was marked as “N/A” (not applicable).

Table 5. File Viewer Instance Chart Showing Fail Results (.doc).

GETDATA EXPLORER VIEW									
DOC	Header/ Footer	Font	Images/ Diagrams	Bullets	Hyperlink	Page Count	Text Search	Print	
1994 to 1999									
File 1	FAIL	FAIL	FAIL	FAIL	N/A	FAIL	PASS	PASS	
File 2	FAIL	FAIL	N/A	FAIL	FAIL	FAIL	PASS	PASS	
File 3	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	PASS	PASS	
2000 to 2005									
File 1	FAIL	PASS	PASS	FAIL	FAIL	FAIL	PASS	PASS	
File 2	FAIL	PASS	PASS	FAIL	FAIL	FAIL	PASS	PASS	
File 3	FAIL	PASS	PASS	FAIL	N/A	FAIL	PASS	PASS	
2006 to 2011									
File 1	FAIL	PASS	PASS	FAIL	FAIL	FAIL	PASS	PASS	
File 2	FAIL	PASS	PASS	FAIL	N/A	FAIL	PASS	PASS	
File 3	FAIL	PASS	N/A	FAIL	FAIL	FAIL	PASS	PASS	

The pass/fails for each file viewer and all 12 file formats were compiled into three charts, showing the performance of the file viewer on the three time blocks: newer files dated from 2006 to 2011; somewhat older files from 2000 to 2005; and older files from 1994 to 1999 (see Tables 6, 7, and 8).

Table 6: File Viewer Capabilities by File Format (2006-2011).

File Viewer	DOC	PDF	PPT	XLS	MSG	HTM	DWG	VSD	JPG	TIF	MOV	AVI
Accessory Software File Viewer	PASS	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	PASS
FileStream Turbo Browser	FAIL	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	PASS	PASS
GetData Explorer View	FAIL	PASS	PASS	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	FAIL
Irfan View	FAIL	FAIL	FAIL	FAIL	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	PASS
Quick View Plus	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	FAIL
UV ViewSoft	PASS	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS

Table 7. File Viewer Capabilities by File Format (2000-2005)

File Viewer	DOC	PDF	PPT	XLS	MS G	HT M	DW G	VSD	JPG	TIF	MO V	AVI
Accessory Software File Viewer	PASS	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
FileStream Turbo Browser	FAIL	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	PASS	N/A
GetData Explorer View	FAIL	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
Irfan View	FAIL	FAIL	FAIL	FAIL	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
Quick View Plus	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	N/A
UV ViewSoft	PASS	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	N/A

Table 8. File Viewer Capabilities by File Format (1994-1999)

File Viewer	DOC	PDF	PPT	XLS	MS G	HT M	DW G	VSD	JPG	TIF	MO V	AVI
Accessory Software File Viewer	PASS	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
FileStream Turbo Browser	FAIL	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	PASS	N/A
GetData Explorer View	FAIL	PASS	FAIL	FAIL	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
Irfan View	FAIL	FAIL	FAIL	FAIL	FAIL	PASS	FAIL	FAIL	PASS	PASS	FAIL	N/A
Quick View Plus	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	N/A
UV ViewSoft	FAIL	PASS	FAIL	PASS	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	N/A

None of the file viewers successfully rendered all 12 file formats. Two file viewers were able to open 10 out of 12 formats: FileStream Turbo Browser and Quick View Plus. While Turbo Browser was unable to open .doc and .vsd files, Quick View was unable to open .mov or .avi files. Interestingly, all file viewers rendered the .jpg and .tif image files, for all three time blocks.

Using the File Viewer Capabilities by File Format charts, a final chart was created, indicating File Viewer Rendering Capabilities by File Format Type (see Table 9). A diagonal bar was used to indicate where the test results did not match expectations.

In terms of results, all file viewers successfully rendered at least one file format category. However, in the context of the Surrey testing and test environment, only Quick View Plus test results matched the expected results from product information. Overall, file viewers were more backward compatible than expected, and, in general, if a file viewer could render a file format, it could render older versions of the file format. (There were only two exceptions: GetData Explorer View, for .ppt in 2000-2005 and 1994-1999 and for .xls in 1994-1999; and UV ViewSoft Viewer, for .doc in 1994-1999.) Image file formats

Table 9. File Viewer Rendering Capabilities by File Format Type (based on testing)

Product	Text	Data	Email	Drawing	Images	Moving Images
Accessory Software File Viewer	No	Yes	No	No	Yes	No
FileStream Turbo Browser	No	Yes	Yes	No	Yes	Yes
GetData Explorer View	No	No	No	No	Yes	No
Irfan View	No	No	No	No	Yes	No
Quick View Plus	Yes	Yes	Yes	Yes	Yes	No
UV ViewSoft	No	Yes	No	No	Yes	Yes

(i.e., .jpg, .tif) were rendered by all six file viewers, while other formats were not rendered by a number of file viewers (i.e., .doc, .ppt, .msg, .dwg, .vsd, .mov).

In terms of results, all file viewers successfully rendered at least one file format category. However, in the context of the Surrey testing and test environment, only Quick View Plus test results matched the expected results from product information. Overall, file viewers were more backward compatible than expected, and, in general, if a file viewer could render a file format, it could render older versions of the file format. (There were only two exceptions: GetData Explorer View, for .ppt in 2000-2005 and 1994-1999 and for .xls in 1994-1999; and UV ViewSoft Viewer, for .doc in 1994-1999.) Image file formats (i.e., .jpg, .tif) were rendered by all six file viewers, while other formats were not rendered by a number of file viewers (i.e., .doc, .ppt, .msg, .dwg, .vsd, .mov).

File viewers also demonstrated the conventional data/content loss limitations commonly noted in .pdf conversions, namely that: formulas were not displayed (.xls); slide presentation and animation was missing (.ppt); and hyperlinks did not work (.doc). This makes sense, as a number of the file viewers used the Adobe Acrobat viewer for on-the-fly conversion as well as .pdf rendering.

In fact, measuring how well the six file viewers rendered files made the co-investigators more aware of .pdf format limitations. There are many benefits to using .pdf as a preservation format: an open standard; a strong working group; a new version in development (i.e., PDF Universal Access, or PDF/UA); a fixed form that is portable, reliable, and interoperable; and billions of instances in existence. However, there are challenges in using the .pdf format for file format conversion. These are challenges are outlined (and debated) in a blog thread entitled, “After Flash, PDF Must Die” (Huber, 2012) and include: a non-reversible transformation requiring the preservation of native files; content/data loss (e.g., formulas, presentation, animation, hyperlinks); and tagging requirements (i.e., to optimize retrieval or re-use across devices). An interesting argument is made that the .pdf format may be “the software version of microfiche,” and that in the future, libraries will need to implement .pdf readers to provide access to the billions of files being created today. Time will tell, although it is interesting to note that the .tif format was seen as a de facto preservation format through the 1990s and early 2000s and now is regularly passed over in favour of the .pdf format. This discussion is continued later on, in the context of non-reversible transformation.

4.4 What Role Might File Viewers Play in Digital Preservation?

File viewers do allow rendering of file formats on-the-fly, with results similar to digital conversion and without the some of the resource requirements, technical difficulties, or migration risks. For the file formats selected in the general study, the file viewers proved to be backward compatible, and able to render files over an 18-year period, without accessing the native applications. File viewers are useful appraisal tools, as files can be rendered without opening native applications which can be difficult to effectively manage during appraisal activities. In some environments, file viewers enable access where native applications that are not resident in the appraisal environment, and also alleviate software licensing costs. For these reasons, file viewers may be considered by some organization to be a viable tool, or even a component of a digital preservation strategy.

File viewers do not overcome problems associated with content/data loss but do underline the somewhat overlooked problem of non-reversible transformation. Although some researchers believe that digital objects “do not need to remain in a state that is unchanged” (Wilson, 2007b, p. 4), researchers on the CAMiLEON project participants noted that, “Existing methods of preserving digital data often fall short of accurately preserving and authentically rendering an original digital document...” and that, “There are many drawbacks with this strategy of ‘traditional migration’... Any errors or omissions from a transformation will propagate...” (Mellor *et al.*, 2002, p. 517). In the CAMiLEON project, “migration on request” was proposed as an alternative strategy to migration conversion. Here, a “digital object is simply archived in its original format,” based on “the principle of always maintaining the original bytestream” (p. 518). The standard for preservation conversion was reversible transformation, as “the only way of ensuring a migration step has been completed without error is by the proof of reversible migration” (p. 519).

The problem with both preservation migration and file viewer conversion is that content is often lost through the representation of the native byte stream in the new format. Through this examination of file viewers, the most important consideration was how to assess the file viewers in terms of properties and characteristics. Depending on the expectations for properties and characteristics, test results would change so that more file viewers might “fail” or “pass.” Although properties, in the sense of file property metadata, are clearly conveyed through standards, data dictionaries, and many other forums, characteristics are more difficult to assess, and further work is likely needed. Based on this study, there are at least three categories of characteristics that are important for assessing file format conversion: structure-related (e.g., cells, line breaks, page breaks, tables, and bullets); appearance-related (e.g., font size and colour, images, and diagrams), and behaviour- related (e.g., formulas, macros, and slide presentation and animations). Similar observations were noted in the *InSPECT Significant Properties Report* (Wilson, 2007b), with reference to content, context, appearance, structure, and behaviour.

6. Conclusion

In closing, the co-investigators recognize the migration-conversion approach as the primary digital preservation strategy in place in archives today. This strategy provides important risk insurance for digital objects, and especially those in danger of immediate obsolescence. For some organizations, the risk of not having electronic information available in an accessible format largely outweighs the total costs of file migration. However, the migration-conversion strategy is not perfect, as characteristics are often lost during file transformations. With many institutions maintaining the native files in addition to a preservation copy, opportunities exist to pursue complementary strategies. For these reasons, the co-

investigators suggest that file viewers provide an opportunity to leverage native files in a digital preservation strategy. Here, the co-investigators note an extensive body of work on file formats in progress beyond the field of archives and records management, and the need to collaborate with these other fields of study, including software development.

Acknowledgements

The co-investigators would like to acknowledge the work of the InterPARES 3 graduate research assistants, Jen Busch and Sergey Kovynev, who participated in the literature review for the *General Study on File Viewers*, the contributions of Dr. Luciana Duranti, InterPARES Project Director, and the input of Team Canada members.

References

- Accessory Software File Viewer 9. Accessed February 23, 2012. <http://www.accessoryware.com/FileView.htm>.
- Amirani, M. C., M. Toorani, and A. A. B. Shirazi. "A New Approach to Content-Based File Type Detection." In *Proceedings of the 13th IEEE Symposium on Computers and Communications (ISCC 2008), July 2008*, 700-705.
- Anderson, P., and T. Teitelbaum. "Software Inspection Using Codesurfer." In *Proceedings of the First Workshop on Inspection in Software Engineering (WISE 2001), Paris. July 2001*, 1-9.
- ARMA International. "The Digital Records Conversion Process: Program Planning, Requirements, Procedures." USA: ARMA International, 2007.
- CNET Download.com. Accessed February 23, 2012. <http://www.download.com>.
- Daeja ViewONE. Accessed February 23, 2012. <http://www.daeja.com/products/viewone-pro-overview.asp>.
- Dhanalakshmi, R., and C. Chellappan. "File Format Identification and Information Extraction." In *Proceedings of the 2009 World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), Coimbatore, India, 9-11 December 2009*, edited by Ajith Abraham, Andre Carvalho, Francisco Herrera and Vijayalakshmi Pai, 1497-1501. 2009. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5393688.
- Enns, L., and G. Badesha. *Shared Drive Migration Toolkit*. Canada: City of Surrey, 2011.
- FileStream Turbo Browser. Accessed February 23, 2012. <http://www.filestream.com/turbobrowser/>.
- The Florida Centre for Library Automation. "Florida Digital Archives." 2009. Accessed February 23, 2012. <http://fclaweb.fcla.edu/FDA>.
- GetData Explorer View. Accessed February 23, 2012. <http://www.explorerview.com/>.
- Grace, S., G. Knight, and L. Montague. *Investigating the Significant Properties of Electronic Content Over Time: Final Report*. United Kingdom: Centre for e-Research, King's College London, 2009. http://ie-repository.jisc.ac.uk/526/1/InSPECT_Final_Report_v1.pdf.
- Huber, Serge. "After Flash, Why PDF Must Die!" *Social Business Expert Blog*. April 12, 2012. Accessed May 25, 2012. <http://www.aiim.org/community/blogs/expert/after-flash-why-pdf-must-die-!>.

- Knight, G. *InSPECT Investigating Significant Properties of Electronic Content*. United Kingdom: Centre for e-Research, King's College London, 2007. <http://www.significantproperties.org.uk/>.
- Knight, G. *Investigating the Significant Properties of Electronic Content Over Time: Framework Report*. United Kingdom: Centre for e-Research, King's College London, 2009. <http://www.significantproperties.org.uk/inspect-framework.html>.
- Library of Congress. "Sustainability of Digital Formats: Planning for Library of Congress Collections." 2010. Accessed February 23, 2012. <http://www.digitalpreservation.gov/formats/>.
- Mancoridis, S., R. C. Holt, and D. A. Penny. "A 'Curriculum-Cycle' Environment For Teaching Programming." In *Proceedings of the 24th SIGCSE Technical Symposium on Computer Science Education (SIGCSE '93)*. New York, 1993.
- McDaniel, M., and M. H. Heydari. "Content Based File Type Detection Algorithms." In *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*. 2002. <http://dl.acm.org/citation.cfm?id=821828>.
- McHenry, K., R. Kooper, and P. Bajcsy. "Towards a Universal, Quantifiable, and Scalable File Format Converter." In *Fifth IEEE International Conference on e-Science*. Oxford, December 9-11, 2009. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5380873.
- Mellor, P., P. Wheatly, and D. Sergeant. "Migration on Request, A Practical Technique for Preservation." In *Research and Advanced Technology for Digital Libraries, 6th European Conference Proceedings (ECDL 2002)*, Rome, 16-18 September 2002, 516-526.
- Quick View Plus. Accessed February 23, 2012. <http://www.avantstar.com/metro/home/Products/QuickViewPlusStandardEdition>.
- Rogers, C., S. Malmas, and L. Enns. "Case Study Final Report: CS 14 City of Surrey Policies, Guidelines and Procedures for a Drive Migration Project as Part of an Enterprise Content Management Program." Vancouver: InterPARES 3 Project, 2011.
- Roseman, M., and S. Greenberg. "Building Real-Time Groupware with GroupKit, A Groupware Toolkit." *ACM Transactions on Computer-Human Interaction* 3, no. 1 (1995): 1-30.
- Skiljan, I. "Irfan View." 2005. Accessed February 23, 2012. <http://www.irfanview.ca/>.
- Sourceforge. Accessed February 23, 2012. <http://sourceforge.net/>.
- Stellant Outside In. Accessed February 23, 2012. <http://www.oracle.com/us/technologies/embedded/025613.htm>.
- Universal Viewer. Accessed February 23, 2012. <http://www.uvviewsoft.com/>.
- Wilson, A. (2007a). "The Significant Properties of Digital Objects." JISC Significant Properties Workshop, British Library, April 2008. <http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops>.
- Wilson, A. (2007b). "InSPECT: Significant Properties Report." *Arts and Humanities Data Service*, 10 April 2007. http://www.significantproperties.org.uk/wp22_significant_properties.pdf.

Autonomic Preservation of “Access Copies” of Digital Contents

Walter Allasia,¹ Fabrizio Falchi,² Francesco Gallo¹ and Carlo Meghini²

¹EURIX S.r.l. walter.allasia@eurixgroup.com; ²ISTI-CNR

Abstract

The paper proposes the abstract approach enabling the delegation of the preservation processes that can be applied automatically to the access copies. The current standard, OAIS, is a reference model only and does not prescribe or even guide implementation. In particular, OAIS is designed around the Archival Information Package, leaving the Dissemination Information Package outside its boundaries. At the same time the “access copies” created as Dissemination Packages, must be preserved as well and very often can be considered as the cultural heritage the user is able to access. Small institutions as well as private users, usually cannot afford a dedicated Archival Information System for preservation and their access copies often constitute the contents themselves. This paper investigates and analyses the problem of autonomic management of the “access copies” preservation. Autonomous agents have been conceptualized in order to manage the stored contents and to keep them updated. The representation of dependencies and obsolescence has been analysed for allowing the automatic application of migration paths.

Authors

Walter Allasia graduated in Physics in 1994 at the University of Turin with a thesis on the design of front-end electronic for nuclear detectors, tested at CERN, joined EURIX in 1997 and worked as software architect and designer of Enterprise Object Oriented Applications. Since 2001 he is Professor by contract at Physics University of Turin. He has been involved in several international projects such as PrestoSpace, PrestoPRIME, SAPIR, DMP and contributed to several MPEG standards such as MPQF, MAF, MPEG-M. During the Ph.D. he was involved in setting up an Interoperable Digital Rights Management Infrastructure on a structured overlay network. Currently his research is focusing on Information Retrieval and Digital Preservation. Since March 2011 he is co-chairing the MPEG Ad Hoc Group on Multimedia Preservation.

Fabrizio Falchi is a researcher at ISTI-CNR in Pisa working on multimedia information retrieval, multimedia content management systems, content-based image retrieval, Peer-to-Peer systems. He received the MSc in Computer Engineering from the University of Pisa, a Ph.D. in Information Engineering from University of Pisa, and a Ph.D. in Informatics from Faculty of Informatics of Masaryk University of Brno. He has a MBA in “Innovation Management & Services Engineering” from Scuola Superiore Sant’Anna in Pisa. He has been involved in several national and international project including SAPIR, VISITO Tuscany, NeP4B, ASSETS.

Francesco Gallo graduated in Physics in 2001 and obtained a Ph.D. in High Energy Physics in 2004 at University of Torino. From 2000 to 2006, as associate member of INFN Torino, he participated to international experiments at Stanford Linear Accelerator Center, California. His research activity focused on the study and development of software algorithms and analysis tools for the precise measurement and validation of the Standard Model of fundamental interactions. Since 2006 he has been involved in several R&D projects in EURIX, such as SAPIR and PrestoPRIME, covering different research areas: information retrieval, digital preservation, digital media processing. Currently he is responsible for the development of the PrestoPRIME preservation platform and is editor of the ISO/IEC 21000-19 (MPEG-21 CEL) standard.

Carlo Meghini is a prime researcher at ISTI, working in the area of digital libraries and digital preservation. In the area of digital libraries, he has been involved in the DELOS Network of Excellence in

Digital Libraries, contributing to the DELOS Reference Model for Digital Libraries; he participated in the FP6 Integrated Project BRICKS, aiming at developing a distributed Digital Library Management System, in the DL.org coordination action, and is involved in the making of Europeana since 2007, through the EDLnet, Europeana version 1.0, Europeana version 2.0 and ASSETS Best Practice Networks. In the area of digital preservation, he has been involved in the CASPAR project, an FP6 Integrated Project aiming at developing an OAIS-based architecture for preservation; he has also taught the OAIS Reference Model in several events organized by the CASPAR Project in conjunction with PLANETS and DPE Network of Excellence in Digital Preservation.

1. Introduction

Setting up an Archival Information System for digital preservation is a complex topic covering several aspects. The paper proposes to make use of Autonomic agents for setting up a preservation framework able to manage the digital “obsolescence” automatically. Focusing the attention to a specific context, the access copies, the authors introduce the backbone for enabling the autonomic preserving. Section 2 reports the state of the art and current technologies in the field of digital preservation and autonomic systems. Section 3 introduces the Autonomic approach discussed in this paper, how to discover obsolescence and how to manage it autonomically, such as for example how to identify automatically the best migration path for contents and formats according to their relationships and dependencies: a model for these descriptions is presented in Section 4. Finally section 5 gives a preliminary design of the autonomic preservation, specifically addressing the “access copies”. Section 6 provides an overview of some future work that can be already foreseen, some currently investigated by the authors.

2. Background

The contribution proposed by this paper is built on top of the OAIS model and the concept of autonomic computing, which are briefly introduced in the following. Open Archival Information System (OAIS) reference model¹ was developed to standardize digital preservation practice and provides a set of recommendations for preservation program implementation. It uses Representation Information to transform a data object into an information object and the Archival Information Package (AIP) to enable collections of information to be preserved over time by identifying the metadata needed for preservation. OAIS model has been adopted in several projects and initiatives as the basis for the design and implementation of digital preservation systems.

Introduced by Paul Horn at IBM in 2001,² Autonomic Computing refers to self-managing characteristics of distributed computing resources, adapting to unpredictable changes while hiding intrinsic complexity to operators and users. In Brazier et al. 2009,³ the relationship between autonomic computing and agents has been studied. In particular, knowledge and reasoning, planning and scheduling,

¹ Consultative Committee for Space Data Systems. *Recommendation for Space Data System Standards: Reference Model for an Open Archival Information System (OAIS). Blue Book*, 2002.

² P. Horn, “Autonomic computing: IBM’s perspective on the state of information technology,” also known as IBM’s Autonomic Computing Manifesto, IBM (October 2001).
http://www.research.ibm.com/autonomic/manifesto/autonomic_computing.pdf.

³ Frances M. T. Brazier, Jeffrey O. Kephart, H. Van Dyke Parunak, and Michael N. Huhns, “Agents and Service-Oriented Computing for Autonomic Computing: A Research Agenda,” *IEEE Internet Computing* 13, no. 3 (May 2009): 82-87.

and inter-agent communication developed for individual agent systems are especially appropriate for autonomic computing.

- Several projects have been funded so far, focusing on the topics mentioned above. An exhaustive list is beyond the scope of the paper, the following projects and initiatives focused on digital preservation using different approaches and are worth mentioning:
- Preservation and Long-Term Access Through Networked Services (Planets)⁴ was a project co-funded by the European Community under FP6 to build practical services and tools to help ensure long-term access to digital cultural and scientific assets.
- Cultural Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR)⁵ was EU FP6 project that aimed at implementing, extending, and validating the OAIS reference model, enhance the techniques for capturing Representation Information and other preservation related information for content objects.
- Automated Obsolescence Notification System projects (AONS and AONSII)⁶ have been software development projects by the National Library of Australia in conjunction with the Australian Partnership for Sustainable Repositories (APR). The projects aimed at developing a software tool that automatically finds and report indicators of obsolescence risks, to help repository managers decide if preservation action is needed. As reported in Pearson et al. 2008,⁷ there is still a mismatch between this objective and the available sources of information on file formats and further development related to format obsolescence is needed.
- SCience Data Infrastructure for Preservation—with focus on Earth Science (SCIPID-es)⁸ is a EU FP7 project that aims at delivering generic services for science data preservation as part of the data infrastructure for e-science and to build on the experience of the ESA Earth Observation Long-term Data Preservation (LTDP) programme.
- Enabling kNowledge Sustainability, Usability and Recovery for Economic value (ENSURE)⁹ is a research project funded by the European Community under FP7 Programme, to extend the state-of-the-art in digital preservation to heterogeneous data. The use cases adopted come from healthcare, clinical trials, and financial services.
- From Collect-All Archives to Community Memories (ARCOMEM)¹⁰ aims at helping to transform archives into collective memories more tightly integrated with their community of

⁴ <http://www.planets-project.eu/>

⁵ <http://www.casparpreserves.eu/>

⁶ J. Curtis, P. Koerbin, P. Raftos, D. Berriman, and J. Hunter, "AONS - An Obsolescence Detection and Notification Service for Web Archives and Digital Repositories," *New Review on Hypermedia and Multimedia* 13, no. 1 (2007): 39-54; David Pearson, "AONS II: continuing the trend towards preservation software "Nirvana"" (paper presented at iPres2007, Beijing, China, October 11-12, 2007).

⁷ D. Pearson and C. Webb, "Defining File Format Obsolescence: A Risky Journey," *International Journal of Digital Curation* 3, no. 1 (2008): 89-106.

⁸ <http://www.scidip-es.eu/>

⁹ <http://ensure-fp7.eu/>

¹⁰ <http://www.arcomem.eu/>

users, exploiting Web 2.0 and the wisdom of crowds to make web archiving a more selective and meaning-based process.

- SCALable Preservation Environments (SCAPE)¹¹ addresses scalability of large-scale digital preservation workflows aiming at enhancing the state-of-the-art developing infrastructure, providing a framework for automated, quality-assured preservation workflows, and integrating these components with a policy-based preservation planning and watch system.
- Digital Preservation for Timeless Business Processes and Services (TIMBUS)¹² explores the digital preservation problem in scenarios in which the important digital information to be preserved is the execution context within data is processed, analysed, transformed and rendered.
- Alliance Permanent Access to the Records of Science in Europe (APARSEN)¹³ is a Network of Excellence looks across the work in digital preservation which is carried out in Europe and tries to bring it together under a common vision.
- PrestoPRIME¹⁴ is a research project funded under FP7 programme aiming at developing solutions for the long-term preservation of audio-visual digital media objects, programmes and collections, and finding ways to increase access by integrating the media archives with European on-line digital libraries in a digital preservation framework. Tools and services will delivered through a networked CompetenceCentre.¹⁵

The new directions in long-term digital preservation as covered by the ENSURE, ARCOMEM, SCAPE and TIMBUS EU projects have been recently discussed in Edelstein et al. 2011.¹⁶ The discussion underlines that ARCOMEM stands alone in dealing with publicly available and non-regulated data while TIMBUS is the only one focusing on the environments. ENSURE and TIMBUS are both motivated in part by accurate risk assessment and preservation lifecycle issues related to regulations and together with SCAPE they also address the scalability of the infrastructures. Central to all of the stated projects is the ability to define what data needs to be preserved. The automation of the preservation lifecycle is being dealt by all the project except ARCOMEM. While SCAPE will be creating preservation lifecycles for deployment on large computational clusters, ENSURE and TIMBUS will examine how to extend existing lifecycle management tools to meet the additional requirements that digital preservation entails. Additionally, projects such as CASPAR and PrestoPRIME investigated the issues related to enabling future access, including representation and management of rights associated to digital content. In particular, PrestoPRIME developed a rights ontology model based on the analysis of narrative contracts in the B2B environment, which is currently under standardization as part 19 of MPEG-21 (ISO/IEC 21000-19). The PrestoCentre, the competence center established by PrestoPRIME, aims at collecting best practices, guidelines and solutions at a European level, including support for standardization activities

¹¹ <http://www.scape-project.eu/>

¹² <http://www.timbusproject.net/>

¹³ <http://www.alliancepermanentaccess.org/>

¹⁴ <http://www.prestoprime.eu>

¹⁵ <http://www.prestocentre.eu>

¹⁶ Orit Edelstein, Michael Factor, Ross King, Thomas Risse, Eliot Salant, and Philip Taylor, "Evolving Domains, Problems and Solutions for Long-term Digital Preservation," In *Proceedings of iPRES 2011 - 8th International Conference on Preservation of Digital Objects*, 2011.

such as MPEG Multimedia Preservation Description Information, aiming at providing a standard interoperable preservation description that is capable of preserving multimedia content for long-term.

3. The Autonomic Approach for Digital Preservation

Starting from the OAIS¹⁷ model and from the experience matured in the projects cited in the previous section, we focus our attention to the access copies, going a little bit beyond the OAIS, introducing the autonomic management of preservation within the functional block “preservation planning”, limited to access copies of the archive. According to Kephart et al. 2003,¹⁸ we can image the “preservation planning” component made up of agents, autonomic agents linked together between federated networks of archives, able to perform two main tasks autonomously:

- discover** the obsolescence of preserved contents
- migrate** the obsolete contents to the most appropriate format

Software already updates itself and the reader can easily recall the automatic software update notification for packages such as those provided by Adobe and OpenOffice. Actually these examples are dealing with specific software installed on specific hardware and last but not least, the software vendor with a specific registry updated for storing and publishing the obsolescence and the “update paths” to be followed.

Actually on the one hand these proprietary software packages are able to self-update, migrating automatically to the newer version, but on the other hand they have no knowledge about dependencies and consequences derived from the proposed update. Many times, the user is suffering problems from the self update of software applications, because the operation breaks dependencies or, worse, update some libraries shared to other software components that will never work again. Summing up, the “*self-update of software packages*” represents a menace for the user instead of being a benefit. Specific aspects to point out are:

- Self-update is usually not able to recognize and identify which other components are making use of the obsolete software, that potentially cannot work anymore. As example, moving from Excel 2007 to Excel 2010 requires a spreadsheet format migration from xsl to xlsx that is not always straightforward and maybe not working especially if file protection is used. Furthermore, if some application is specifically written for having “xls” files as input, it will never work again after the update.
- Self-update is not able to evaluate which software components have been used in the older versions, that are candidate to be broken and do not work anymore.

Hence, instead of simply asking the user “A new version of software X is available, do you want to install it?” it would be better if the self-update process was able to ask the following question: “A new version of software X is available, if you want to install it you have to migrate the following packages and in order to have these software component still running, you have to move these libraries to something else, etc....” Unfortunately as the reader has already experienced, the self-update of software packages is a serious

¹⁷ OAIS 2002.

¹⁸ Jeffrey O. Kephart and David M. Chess, “The Vision of Autonomic Computing,” *IBM Thomas J. Watson Research Center* 0018-9162/03 © 2003 IEEE.

issue and usually completely left to the user's responsibility, without any support and advice on potential damages that the process can cause.

In our case, the autonomic preservation of access copies, the problem is really much wider, without any well established, trusted and complete registry, managing and preserving not only a specific software version but also and mainly the contents and the technology and software components linked or with dependencies associated.

The automatic update of archived master copies is weakly perceived because the responsible archives are already setting up procedures and resources dedicated. Completely different situation is what we have about the access copies, for usually there are no many resources to allocate and an automatic support is welcome and needed.

Even if in Section 2 we have pointed out some experiences and projects addressing these tasks, anyone has come up with a complete solution. Actually most of them are addressing the general preservation issues, hence the loss of a single bit is not allowed for archived copies, resulting in the impossibility to trust any automatic system performing format migrations and conversions without the supervision of human beings. On the other hand, limiting the scope to the "access copies" (or "wife copies"), we can afford the loss of contents as well as the migration to a format chosen automatically by the agent that may be not the best. These faults can be accepted because there is always the master copies and therefore it is possible to re-create access copies at any time, thereby exposing again the original contents.

Moreover, focusing the attention to the "access copies" and to the "access" aspects, we can address also a further issue, going beyond OAIS, the preservation to the "publication system".

The "accessibility" to digital contents is not only limited to the contents but also the system providing the contents as well, connected to the clients with several different technologies involved.

We can figure out an "autonomic manager" responsible for the managed elements that are made up of our "access copies" together with the systems related to the access process (publication system, web servers, etc.). The "autonomic manager" is responsible for performing the following functions without the human intervention:¹⁹ self-configuration, self-healing, self-optimization, self-protection.

In our "preservation" context these functions can be implemented as follows:

Self-configuration means have a complete knowledge of preserved formats, especially if preserved contents are multimedia where we have wrappers/containers and different tracks for audio/video/text with associated codecs. Also the knowledge of all the "accessibility" aspects in charge to a specific node.

Monitor must also check what is going to be performed within the federated networks of other "autonomic agents", selected choices and migrations strategies applied. More specifically, in this phase the autonomic manager will evaluate configuration files, registries, makefiles, classpaths, shared and linked dynamic libraries, etc., generally every software and hardware structure having "usage" dependencies each other and within the architectural components. Agent must configure itself in order to receive the obsolescence events and selected the best actions to perform.

Section 4 describes how to represent dependencies in order to set up the needed framework that the agent can apply in an autonomous way.

¹⁹ Brazier et al. 2009.

Self-protection means that the manager must proactively identify preservation issues and take appropriate actions. The “agent” receive messages about obsolescence risks and issues to be managed. Hence self-protection means that he will take the most appropriate actions in order to remove the issues.

Self-optimization means that once the autonomic manager has analysed a specific issue and decided to intervene for self-protection as well as self-healing, a ranking of the different, possible plans will be derived, taking into account efficiency and performance requirements, constraints and the rules to be followed. The ranking will be used to select the most convenient plan, either autonomously or by asking human intervention.

Self-healing is the action of autonomously evaluate the knowledge acquired in the monitor activity, in order to come up with suggestions and decisions to implement. This is the most important activity delegated to a preservation system: the healing analysis must decide the most appropriate action to be taken for preserving the access copies and the access system at risk. The following Section 4 describes how to represent the obsolescences in order to apply the rules that the agent uses for self-healing in an autonomous way from obsolescence. Self-healing represents the “cure” to the “obsolescence disease”.

Having in mind the above concepts and taking into account next section (Section 4) describing the representation of obsolescence and dependencies, we can sum up and design the overall system (Section 5) able to autonomously manage the preservation of access copies.

4. A Model for Dependencies and Obsolescence Representation

Work on digital preservation so far has been largely inspired by the OAIS reference model,²⁰ which defines the information and the functional model of an archival system for the preservation of some information content. Although OAIS was never meant to be a design pattern, and even less a functional specification, it has been essentially interpreted that way. As a result, projects addressing digital preservation (from now on, DP for short) almost invariably end up designing *ad hoc* DP systems whose information structures and functional architectures reflect, respectively, the information and the functional models of OAIS. These DP systems are used as external entities, which are handed over the output of some other system, in order to preserve such output. In these approaches, preservation is achieved through a de-contextualization of the information to be preserved, and a great effort is required in order to endow the de-contextualized information with enough knowledge (Representation Information and PDI) to allow its interpretability in the future.

Contrary to these approaches, we think that the long-term preservation of information should be the effect of a quality that any information system can exhibit. We call such quality as longevity. Longevity is the ability of an information system of achieving a set of goals that have a duration in time, no matter how distant in the future. Whenever the duration in time of a goal extends beyond the lifetime of the technology used to achieve that goal, a DP problem arises. Thus, longevity captures DP as it is usually defined.

²⁰ OAIS 2002.

Current software development methodologies take a great care to guarantee that newly constructed information systems (from now on ISs, for short) satisfy a set of goals at IS release time. To this end, these methodologies provide languages for explicitly representing goals as requirements, and for relating requirements to the artefacts that document the software development process, from the design of the IS down to its final architecture. In the UML,²¹ for instance, requirements are represented as use cases, and are related to the structural and behavioural elements that implement them. This relation is a realization: it associates use cases to collaborations, defined as “societies of classes, interfaces, and other elements that work together to provide some cooperative behaviour.”²² Interfaces, in turn, are realized by components. Components are the basic elements of architectures, defined as “physical and replaceable parts of a system that conforms to and realizes a set of interfaces.”²³ By composing the realization relation, then, the UML allows to build a chain that connects requirements to the interfaces that realize them, and interfaces to the components that realize them. This chain gives therefore a very solid representation of the connection between the goals of an IS and the physical parts of the IS that achieve these goals, and as such it is crucial for longevity.

Unfortunately, OAIS has driven attention away from this chain, by proposing an Information Model that does not include any element of the chain. However, we argue that this chain is crucial for achieving a core functionality of OAIS, namely the “Develop Preservation Strategies and Standards” function, which is part of the Preservation Planning functional entity of the OAIS Functional Model. In fact, our proposal can be understood as directed towards supporting the development of preservation strategies through automation.

Any IS developed according to a principled methodology is thus born to potentially satisfy longevity. But in order to effectively satisfy longevity, an IS must successfully cope with the passage of time. The passage of time determines the obsolescence of the three fundamental elements of an IS:

- *Software components*: The obsolescence of software components is a well-known fact. It is caused by the perpetual evolution of technology and affects longevity due to the fact that the interfaces of obsolescent components will no longer be implemented, causing a set of goals to be no longer achieved.
- *Contents*: The obsolescence of the contents of an IS (understood as multimedia complex objects) is mostly caused by the degradation of media. As a result, some component using the content stored on the obsolescent media is no longer able to function, again causing a set of goals to be no longer achieved.
- *Metadata*: the obsolescence of metadata is due to the fact that the ontologies evolve: Some term in an ontology may fall out of usage, or may change its meaning, and new terms may come into usage. As a consequence, the metadata that use the out-dated or the changed terms, or that do not use the newly introduced terms, do not convey the same meaning as before to their users (be they software or humans in the designated community), and as such they may no longer be used for the same purpose.

²¹ Unified Modeling Language™ (UML®) - Version (2.4.1) has been formally published by ISO as the 2012 edition standard: ISO/IEC 19505-1 and 19505-2. <http://www.omg.org/spec/UML/ISO/19505-2/PDF>.

²² Ibid., p. 371.

²³ Ibid., p. 345.

In order to cope with obsolescence, the architecture on an IS needs to evolve. Evolving architectures is at the heart of digital preservation. It amounts to replace a set of components of the architecture, whether software, contents or metadata, with different ones so that a new architecture is obtained *that achieves all goals in place*. In the case of software components, evolution means replacement with one or more different software components in order to preserve functionality; in the case of contents, evolution means transformation from one format to another one in order to preserve information; in the case of metadata, evolution means re-writing according to a different ontology, in order to preserve meaning. The evolution of the architecture of an IS is intended to capture all these three different kinds of evolutions.

In this paper we do not consider the evolution of metadata, which requires techniques that significantly differ from those required for the evolution of software components and contents.

Most of the times the replacement of a component has consequences that impact other components of the architecture. Sometimes there is no unique replacement to be made, and a decision amongst a set of alternatives has to be taken, based on some utility function. Sometimes, it is not possible to know a priori whether an evolution of the current architecture satisfies all requirements in place, and a certain number of tests have to be executed to validate the evolution.

4.1 The Approach

The methodology that we propose is based on the principle that the IS is endowed with, and uses an explicit representation of its own goals, of its own functional architecture, and of the relationships between the former and the latter that state how the goals are realized by the functions provided by the architecture. As already argued above, languages for expressing requirements and functional architectures already exist, and are currently used for software development. Those languages need to be extended by combining them with languages for representing goals, which have been researched in the last decade and are now at a mature stage of development.²⁴ Indeed, goals stand at a more abstract level than requirements; moreover, they have their own structure and have a temporal dimension whose capturing is crucial for modelling and maintaining longevity. The methodology will also provide the machinery for representing the obsolescence of the IS architecture, as an event situated in time.

Depending on the type of the obsolescence, the corresponding event will carry contextual knowledge that will inform the ensuing evolution process. The methodology will finally provide the algorithms for analysing the obsolescence, determine which goals are affected by it, and propose a ranked set of alternative architectures, each representing a different way of evolving the current architecture in order to attain longevity of the IS. In order to compute alternative architectures, the methodology will need to acquire information about existing components for replacing the obsolescent ones.

Typically, an architecture can be evolved in many different ways, leading to a very large problem space. The methodology will deal with the possible combinatorial explosion by working on two different aspects: from the one hand, it will use a natural ordering amongst architectures ruling out architectures that are *a priori* sub-optimal, i.e., architectures that are functionally redundant, supersets of other architectures, and the like; from the other hand, it will rank architectures based on the combination of two main factors:

²⁴ Unified Modeling Language.

- An application-independent factor, measuring the *distance* between the current architecture and the alternative architecture. Research work is needed to analyse several distance measures in order to determine in an empirical way the one that best fits the use cases at hand.
- An application-dependent factor, measuring the cost of migrating to the new architecture from a domain specific point of view. The identification of application-dependent factors, requires a substantial investment in the analysis of specific domains. Since different domains will have different factors, the application-dependent function is best considered as an input to the methodology.

The final selection of the architecture will be carried out by a human user with the support of the methodology, which will provide this user with the information generated during the architecture ranking process.

5. Designing the Autonomic Preservation of Access Copies

We are integrating the basic concepts of Autonomic Computing, Multi Agent Systems (MAS) and Service Oriented Computing (SOC)²⁵ in order to design a computing preservation system able to manage itself given high level preservation rules provided by the archival administrator. Actually our autonomic system is made up of individual preservation agents, responsible for managing several preservation aspects, most important being: knowledge, reasoning, planning and scheduling.

Knowledge, is related to the “self-configuration” autonomic function (Section 3), where the agents acquire the competence on preserved contents, the associated metadata and the overall Information System (Section 4) in order to identify every link and dependency. It also takes into account rules set up by the archive administrator and the information coming from the connected “preservation agents” (or, as described later, from the “shared environment”) in order to have a complete awareness of the monitored context. **Reasoning** is mainly related to the “self-healing” autonomic function the focus point of the preservation agent, implementing the capability to gather information from the knowledge in order to come up with appropriate preservation actions that the **Planning** will analyse and simulate in the preservation agent environment in order to be able to choose the correct procedures for **Scheduling** the list of tasks to be performed. As already pointed out, Reasoning is the core aspect of our preservation agent. The considered approach is the Stigmergy model, first introduced in,²⁶ where preservation agents collaborate not by direct message exchanged but by jointly making and sensing changes to their shared environments and knowledge.²⁷ In this way we are removing the need of a centralized registry of obsolescence (such as for formats) as we may have been forced to plan (especially in a SOC and SOA approach), moving the registry concept to the complete (and more comprehensive) view of the overall

²⁵ Brazier et al. 2009.

²⁶ P. P. Grassé, “La reconstruction du nid et les Coordinations Inter-Individuelles chez *Bellicositermes Natalensis* et *Cubitermes* sp.” *Insectes Sociaux* 6, no. 1 (1959): 41-84.

²⁷ Brazier et al. 2009; Ozalp Babaoglu, Geoffrey Canright, Andreas Deutsch, Gianni A. Di Caro, Frederick Ducatelle, Luca M. Gambardella, Niloy Ganguly, Márk Jelasity, Roberto Montemanni, Alberto Montresor, and Tore Urnes, “Design patterns from biology for distributed computing,” *ACM Transactions on Autonomous Adaptive Systems* 1, no. 1 (September 2006): 26-66; Blesson Varghese, and Gerard T. McKee, “Applying Autonomic Computing Concepts to Parallel Computing using Intelligent Agents,” *World Academy of Science, Engineering and Technology* 31 (2009): 362-366.

archival system, made up of Information Systems (software, hardware components), metadata and obviously “contents”, representing the knowledge of the preservation agent. A typical scenario could be as follows:

Preservation agent A discovers a weak link in its knowledge, maybe a content format obsolescence, or an application for rendering a specific codec no more available or supported.

- preservation A reasoning, according to its rules, either:
 - Suggests to migrate to a new format selected from an available list or based on a pre-existing knowledge; or
 - Notifies the archive administrator in order to choose the most appropriate migration path to follow.

In the latter case the new rule is added to the knowledge of preservation agent A. Hence the preservation actions are applied to the contents/systems managed by preservation agent A (planning and execution). According to the Stigmergy model,²⁸ preservation agent B is looking at the knowledge of its linked agents, including agent A. New rules are hence available addressing a specific format obsolescence and preservation agent B will reason on contents and systems which is responsible for. In the case B is suffering the same or a partial obsolescence issue, the preservation agent B will autonomously take the appropriate preservation action in order to fix it. It can decide to adopt a partial migration path in order to tailor the preservation action to its specific needs, rules and context. Again, a further preservation agent C in the network of A and B can watch what’s happening and, even if it has no contents under obsolescence risk, he can update its knowledge for future use improving its reasoning capabilities addressing the overall migration path as well as sub-parts and specific weak links management.

The above scenario demonstrates that if on the one hand it is not possible to create a generalized obsolescence registry (as demonstrated by many initiatives (Section 2), on the other hand, making use of autonomic approach with Stigmergy preservation agents, it is possible to create customized registries (tailored specifically to preservation archives), that are distributed according to a shared knowledge managed by preservation agents.

6. Conclusions and Future Work

This paper has introduced a candidate solution for the preservation of digital contents, specifically addressing the “access copies”. The approach has analysed and designed an autonomous model where autonomic systems are able to perform the preservation actions needed to keep contents (and their accessibility) up to date. A method for describing obsolescence has been presented and a solution based on stigmergy approach has been proposed. This position paper opens novel paths to follow enabling automatic and autonomic preservation.

In order to have a shared environment describing the knowledge of “autonomic agents” distributed over the federated networks, we can image to set up services published by Linked Open Data protocols. Making use of “graph” queries it could be possible to extract complex information, links and dependencies, that will be the knowledge base evaluated and modified in order to preserve the access

²⁸ Leslie Marsh and Christian Onof, “Stigmergic epistemology, stigmergic cognition,” *Cognitive Systems Research* 9, no. 1-2 (March 2008): 136-149.

copies. A software model based on the proposed approach can be designed and simulated on open source MAS (Multi Agent Systems) platforms such as Jade²⁹ and Janus,³⁰ adding the needed graphs (or *RDF*³¹ *like*) for representing the knowledge base, continuously monitored and modified by the preservation agents. Currently the authors are investigating artificial life design patterns in order to select the most appropriate stigmergy model and set up a simulator for the experimentation of autonomic preservation processes.

Acknowledgements

This work was partially supported by the PrestoPRIME project, funded by the European Commission under ICT FP7 (Seventh Framework Programme, Contract No. 231161) and the Europeana v2.0, CIP-Thematic Network (contract No: 270902).

²⁹ <http://jade.tilab.com/>

³⁰ <http://www.janus-project.org/Home>

³¹ Resource Description Framework - <http://www.w3.org/RDF/>

A Methodology Framework to Ensure Preservation

Bias and Balance in the Preservation of Digital Heritage

Anca Claudia Prodan

International Graduate School Heritage Studies at Cottbus University, Brandenburg University of Technology

Abstract

This essay draws attention to the cultural challenges of digital preservation. Based on the concepts of “medium bias” and “media balance” coined by Harold Innis a methodology is developed to analyse the influence of digital technology on culture. The results of this analysis are applied to the field of digital preservation and the potential impacts on cultural diversity are discussed. The analysis reveals weaknesses of digital preservation that could hinder a balanced representation of cultures. It also provides suggestions to make digital preservation more inclusive and sensitive to cultural diversity.

Author

Anca Claudia Prodan is a Ph.D. candidate at the International Graduate School Heritage Studies at Cottbus University, Brandenburg University of Technology, Germany. Prior to undertaking her doctoral research, she studied and obtained a bachelor's degree in Anthropology-Philosophy and a master's degree in World Heritage Studies. Her research interests include theories of culture and heritage, cultural policies, and interpersonal and intercultural communication. Her current research is on the Memory of the World Programme and the conceptual changes triggered by digital technologies in this field.

1. Introduction

Today the biggest enemy of digital (born-digital and digitized) documentary heritage is perhaps the process of technological obsolescence. This is reflected in the concerns of libraries and archives where the main aim is to find methods to ensure the long-term survival of digital documents. Currently the discourse on the preservation of digital documentary is dominated by a technical rationality and efforts and resources are concentrated on the challenges that must be overcome to ensure the long-term survival of digital heritage. The preservation community largely acknowledges that digital preservation requires active rather than passive approaches such as benign neglect, and in the case of digital documents, it has turned from conservation of carriers into management of technological change and of the risks that this triggers. Scholars and practitioners also emphasize that in addition to managing technical issues, ensuring the long-term preservation of digital heritage also depends on economic, legal, political, organizational or societal factors.¹ The need to manage cultural challenges is addressed in these accounts to a less extent. My essay is an attempt to draw attention to the complex link between culture and digital preservation and I argue that the preservation of digital heritage should also be concerned about cultural challenges that could hinder digital access on the one hand, and that could result from digital access on the other hand. First I introduce a literature review that relates the topic of digital preservation to the protection of cultural diversity and I motivate the need to consider the cultural dimensions of digital preservation. Then I introduce a theory that gives a more informed understanding of the interplay between culture and

¹ Kevin Bradley, “Defining Digital Sustainability,” in *Preserving Cultural Heritage*, ed Michèle V. Cloonan and Ross Harvey (Illinois: John Hopkins University Press, 2007), 148-163, p. 151.

technology. I use this theory to construct a methodological framework and I apply it to assess the cultural bias of digital technology. I further relate the results of this analysis with the topic of digital preservation to discuss its relevance in the context of cultural diversity. In conclusion I provide a summary of the main points raised throughout the essay.

2. Culture and the Preservation of Digital Heritage

Structurally the Memory of the World Programme (MoW) and the Charter on the Preservation of Digital Heritage—simply the Charter—which arose from the context of MoW, belong to UNESCO's Communication and Information sector. Yet apart from this, the preservation of documentary and digital heritage extends UNESCO's efforts for heritage protection and these concepts are at the same time part of UNESCO's body of heritage concepts that together form the so called common heritage of humanity. From this point of view it is important to emphasize that although preservation of documents means to a great extent preservation of information, UNESCO's standard-setting instruments such as MoW and the Charter advance an understanding of information as heritage. This means that documents addressed by these standard-setting instruments are raised above their informational level and are given ethical dimensions that arise from UNESCO's overall ethical mission. Since MoW and the Charter cannot be separated from UNESCO's mission, its efforts for digital documentary heritage preservation cannot be simply about preservation of information. If this were the case, then these policy instruments would only duplicate the activities of libraries and archives, and this is something that at least the initiators of MoW wanted to avoid from the very beginning.² The Charter, which is an extension of MoW and declares the digital heritage as common heritage, follows the same line of reasoning; thus, the preservation of digital heritage should be considered as part of the broader philosophy of UNESCO. This philosophy, and by extension all standard-setting instruments for heritage, rest on general principles such as human rights, and contribute to overarching objectives such as achieving human development, protecting cultural diversity, or constructing knowledge societies. While acknowledging that in practice these objectives are related and cannot be separated from each other, for analytical purposes I resume my discussion to the notion of cultural diversity. UNESCO's concept of culture has changed throughout its history and an important shift that occurred was passing from a humanistic conception of culture to an anthropological understanding. The concept of culture which is used today was set down in 1982 in the Mexico City Declaration and it reads: "Culture may now be said to be the whole complex of distinctive spiritual, material, intellectual and emotional features that characterize a society or social group. It includes not only the arts and letters, but also modes of life, the fundamental rights of the human being, value systems, traditions and beliefs."³ This definition lies at the basis of the definition of cultural diversity set down in the 2001 Universal Declaration on Cultural Diversity and which says "culture takes diverse forms across time and space. This diversity is embodied in the uniqueness and plurality of the identities of the groups and societies making up humankind."⁴ In my essay I am following this understanding of culture and of

² Abdelaziz Abid, "Memory of the World," in *Report of the First Meeting of the Sub-Committee on Technology* (Paris: UNESCO, 1994).

³ UNESCO, *Mexico City Declaration on Cultural Policies*, World Conference on Cultural Policies, Mexico City, July 26–August 6 (Paris: UNESCO, 1982).

⁴ UNESCO, *Universal Declaration on Cultural Diversity* (Paris: UNESCO, 2001).

cultural diversity. The only difference is that instead of using “groups and societies” I use the notion of culture in the plural, so cultures.

It is worth noting that cultural diversity was initially explicitly mentioned in the Charter. The title of an article in a preliminary draft was “Cultural Diversity and Pluralism”. In a later draft from March 2003 a reference to the 2001 Universal Declaration on Cultural Diversity was incorporated in the preamble of the Charter and the title of article 9 was “Promoting Cultural Diversity”. This draft appears in the Guidelines for the Preservation of Digital Heritage.⁵ Yet in the Charter that was adopted by the General Conference of UNESCO in October 2003 the reference to cultural diversity did not appear anymore, neither as part of the preamble nor as title of article 9, which changed to “Preserving Cultural Heritage”. Why this happened cannot be inferred from drafting documents and discussions that are available for consultation. Nevertheless, the idea was not abandoned and it is reflected in article 9 of the adopted Charter, which reads: “The digital heritage of all regions, countries and communities should be preserved and made accessible, so as to assure over time representation of all peoples, nations, cultures and languages.”⁶ Relating digital preservation with cultural diversity is not only consistent with the Charter but it is perhaps also necessary. In the absence of relating digital preservation to this broader objective one runs the risk of separating digital preservation from context and understanding it as a purely technical process. That this can happen is somehow evident in that the notion of access is approached mainly from a technical perspective. For sure maintaining access from a technical perspective is an important dimension of preservation. Access is said to be the end purpose of preservation because without access preservation doesn’t make sense and this is definitely true. But this is only part of the story and it says nothing about the fact that access at its turn is not an end in itself but a means to a higher goal, be it development, diversity, or dialogue. Thinking about the end purpose of access automatically leads us from thinking about technology to thinking about people. So while agreeing that without access preservation doesn’t make sense, it is important to acknowledge that access makes sense only if it follows considerations about the people for which it is intended. In fact, libraries and archives do think about people when they have in mind the community of users, which informs selection of materials.⁷ At a conceptual level the community of users for which preservation is intended is usually future generations and practitioners argue that a balance between current and future uses should be found.⁸ Future generations is nevertheless too broad a concept to be of practical value and it doesn’t necessarily guide the selection of materials, although it informs the aim of maintaining resources accessible in the long-term. At a practical level libraries and archives have always served a smaller community, usually the community where the institutions are physically located. But digital technology exceeding physical borders enlarges the user community, making its definition very difficult. This becomes even more difficult if one considers that user communities are not static but constantly changing and that a vision about how they change should be incorporated in policies.⁹ And it becomes still more difficult if the

⁵ National Library of Australia, *Guidelines for the Preservation of Digital Heritage* (Paris: UNESCO, 2003), p. 14, accessed February 24, 2010, <http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>.

⁶ UNESCO, *Charter on the Preservation of Digital Heritage* (Paris: UNESCO, 2003).

⁷ Consultative Committee for Space Data Systems (CCSDS), *Reference Model for an Open Archival Information System*, CCSDS 650.0-M-2, CCSDS (Secretariat: Washington, DC, 2012). <http://public.ccsds.org/publications/archive/650x0m2.pdf>.

⁸ John Feather, “Introduction: Principles and Policies,” in *Managing Preservation for Libraries and Archives: Current Practice and Future Developments*, ed. John Feather (England, USA: Ashgate, 2004), 1-26, pp. 7-8.

⁹ CCSDS, *Reference Model*.

purpose is to serve cultural diversity, especially if we consider that the continuous dissemination of computers leads to a more diversified user community, including people from least developed countries, women, elders and other minority groups that so far have not been well represented in the digital environment.

Considerations of diversity in the field of digital preservation do exist, and although they are more modest they deserve a closer attention. Regarding the fact that English is the mainstream language on the Internet, some authors argue on the example of Finland that “for smaller countries representing smaller language groups the dilemma is that the general public, academic institutions, researchers and business might turn to using the English collections on the Web as their main source. This could in the long run diminish the knowledge, research and interest in the national culture as part of Europe’s cultural richness as a whole.”¹⁰ The authors suggest, “the selection of material for the Finnish collections should support the technical solutions for the multilingual use of European collections.”¹¹ Multilingualism is a positive solution and the World Digital Library can be given as example of an international initiative that provides content in various languages. Diversity can also address the subjects or the types of documents preserved, not just the languages. Staying with the example from Finland, the authors also says that an example of diversity “would be digitization of the nineteenth century, which consists of many genres of material such as newspapers, journals, manuscripts, maps, photographs, art and so on. These are the items and genres forming the core collections, in such disciplines as science, history, mathematics, geography, ways of life, education, culture and so on. Each institution is contributing to this pattern, which in the end forms a patchwork of the cultural heritage and extends to other centuries and to international cooperation.”¹²

All these represent positive developments and should be welcomed. Yet, I believe that addressing cultural diversity in a digital environment means more than diversity of contents or languages. The paradigm for digital preservation most often assumes that the content is most important and that the carrier matters only to the extent that it helps transferring the content.¹³ But this runs counter to scientific research that shows how carrier structures content; how it influences people’s understanding of it and impacts their knowledge. So in addition to considering the contents also the technical properties of the carrier should be selected carefully if the aim is to design inclusive policies that in time lead to equal representation of the cultures of the world. Failing to address this could lead to an erosion of diversity, rather than a representation of it. Thinking about diversity and digital preservation requires asking questions about the interplay between culture and technology. For this reason in the chapter below I introduce a theory that re-focuses attention on the materiality of digital information. Later this will enable a discussion of the link between cultural diversity and digital preservation.

3. Theoretical Framework

One theory that studies the interplay between culture and communication technology can be borrowed from the Canadian political economist and communications scholar Harold Adams Innis (1894 – 1952).

¹⁰ Majlis Bremer-Laamanen and Jani Stenvall, “Selection for Digital Preservation: dilemmas and issues,” in *Managing Preservation for Libraries and Archives: Current Practice and Future Developments*, ed. John Feather (England, USA: Ashgate, 2004), 53-65, p. 54.

¹¹ Ibid.

¹² Ibid., p. 64.

¹³ UNESCO, *Memory of the World Register Companion* (Paris: UNESCO, 2011), p. 3, accessed October 23, 2011, <http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/Register%20Companion.pdf>.

Innis passed away before digital technology was developed but his theory is still relevant today and it offers many insights into contemporary subjects as shown by the fact that more and more scholars apply it to various themes. He considered that “a medium of communication has an important influence on the dissemination of knowledge over space and over time and it becomes necessary to study its characteristics in order to appraise its influence in its cultural setting.”¹⁴ Therefore Innis developed a theory that analyses the interrelationships between a medium of communication and cultural change and by tracing such interrelationships over thousands of years of human history he showed how communication media have always been involved in the shaping of cultures, their stability, as well as the structuring of knowledge available to them.¹⁵ He didn’t do this in a deterministic manner. Rather he revealed how the medium shaped culture and was shaped by culture at its turn. His theory was very rich in concepts but for the purpose of my analysis I only borrow two of them: the concepts of bias and balance.

Bias was a generic term he used in order to talk about the characteristics of communication media and their impacts, and he broadly divided them into space-biased media and time-biased media. In the context of documentary heritage preservation medium refers to the physical carrier on which information or content resides. For Innis, however, speech was also a communication medium, even if the carrier was absent; or rather people themselves represented it. He related the notion of bias with space and time because he considered that “the relative emphasis on time or space will imply a bias of significance to the culture in which it is imbedded.”¹⁶ Thus, as Richard Noble explains, “time-biased civilizations tend towards institutional decentralization, an emphasis on the sacred, and efficiency at solving problems of continuity. Their instability arises from their inability to solve problems of space. Space biased civilizations, in contrast; emphasize institutional centralization, imperialism, and efficiency at solving problems of space. Their instability arises from their neglect of the problems of time. Balance is the key to stability.”¹⁷ Interpretations of his discussion of space and time biased media are often too simplistic, emphasizing that this division should be understood “as related to the ability of the message to survive transmission and have impact over space or over time.”¹⁸ Such an interpretation is perhaps derived from Innis’ own writings because he was the one to explain that, “according to its characteristics it [the medium] may be better suited to the dissemination of knowledge over time than over space, particularly if the medium is heavy and durable and not suited to transportation, or to the dissemination of knowledge over space than over time, particularly if the medium is light and easily transported.”¹⁹ Despite this statement he did not mean to simply imply that media were suitable for transmitting messages across time and space but that they predisposed cultures to value expansion or duration. For Heyer this was the result of the fact that “Innis sees the history of civilizations in terms of centripetal and centrifugal forces; in other words, those that aggregate and those that disperse the power they have.”²⁰ Furthermore, a simplistic understanding of the space and time bias creates the impression that the bias of a medium was related to physical matter only but this was hardly the case. Both the concept of bias and his

¹⁴ Innis, *The Bias of Communication* (Toronto: University of Toronto Press, 1995), 33.

¹⁵ Innis, *Empire and Communications* (Toronto: Press Porcépic Limited, 1986).

¹⁶ Innis, *The Bias of Communication*, p. 33.

¹⁷ Richard Noble, “Innis’s Conception of Freedom,” in *Harold Innis in the New Century: Reflections and Refrlections*, ed. Charles R. Acland and William J. Buxton (Québec: McGill-Queen’s University Press, 1999), 31-45, pp. 34-35.

¹⁸ Marshall Soules, “Harold Adams Innis: The Bias of Communications and Monopolies of Power,” 2007, accessed April 9, 2011, <http://www.media-studies.ca/articles/innis.htm>.

¹⁹ Innis, *The Bias of Communication*, p. 33.

²⁰ Paul Heyer, *Harold Innis* (Lanham, Maryland: Rowman and Littlefield Publishers, 2003), 46.

understanding of medium were more complex. According to Heyer when Innis employs “the term medium of communication, it usually does not mean only the raw material used—stone, clay, parchment, or paper—but also the form of communication embodied in that medium—hieroglyphics, cuneiform, or alphabetic writing.”²¹ As for the notion of bias, this was one of the richest in meaning. Van Loon distinguishes in Innis’ writings four dimensions of bias: (1) a bias of matter, relating to physical materials such as stone, paper, electronic wires, microprocessors; (2) a bias of form, relating to how matter is structured and organized; (3) a bias of use, relating to how media are used in social practices; (4) and a bias of know-how, relating to the skills and functions needed to produce outcomes.²² For van Loon it is wrong to interpret Innis’ concept of bias solely from the perspective of the ‘matter’ of communication, because “bias highlights that media-technology is constituted by an interplay between the technological artefact (the tool or better ‘matter’ and ‘form’), its practical applications (usage), as well as the knowledge and skills that are necessary to make it work (know-how).”²³ To this one could add that a medium took its bias also from the context in which it was developed or into which it was introduced. All these factors and their intersection with a medium conditioned the adoption and use of a medium, and consequently its potential impact. So in order to understand the medium bias and its impact it is necessary to analyse its intersections with such factors.

Innis, however, did not carry out an analysis of media just for the sake of determining their bias and this leads to the concept of balance, which is the second concept I borrow from Innis. He was interested in the space and time bias of the media because he believed that the flourishing of human societies depended on ensuring a balance between the concepts of space and time. Should any of the two media become dominant to the point that a monopoly of knowledge is formed, so is the balance disturbed and at several points in history this caused the disintegration of societies.²⁴ His analysis of history and communication media made him conclude that a society lacks stability unless it is capable of dealing with both time and space successfully.²⁵ However, according to Frost there was more at play behind his analysis and she considers that he was informed by a concern for human freedom, for cultural flexibility and for the longevity of cultures. The background of these concerns was the need to return to the human scale and this was reflected in his search for the balance between space and time biased media.²⁶

4. The Bias of Digital Technology

In this chapter the intention is to provide a methodological framework constructed around the notion of bias as described above but before doing this there is need for one clarification. Innis developed a historical analysis; so he looked at things that happened. Yet I agree with Frost that using his theory is useful also for making predictions. Since digital preservation means management of change, risk assessment has become part of it and this requires making predictions about the potential outcomes of change. Innis gave enough examples and covered enough historical periods and this allows us to

²¹ Ibid., p. 63.

²² Joost van Loon, *Media Technology: critical perspectives* (Maidenhead, England: McGraw Hill, New York: Open University Press, 2008), p. 24.

²³ Ibid., p. 26.

²⁴ Innis, *Empire and Communications*.

²⁵ Ibid.

²⁶ Catherine Frost, “How Prometheus is Bound: Applying the Innis Method of Communication Analysis to the Internet,” *Canadian Journal of Communication* 28, no. 1 (2003): 9-24.

understand the steps one can take for assessing the impacts, which follow the introduction of a medium and its use. Different authors used his method in different ways, depending on the subject they applied it to and the concepts they borrowed.²⁷ In this essay I am following the method suggested by Catherine Frost, which implies taking three steps: assessing the pre-existing conditions which surrounded the roots of a medium; analysing the characteristics of the medium; and identifying the phenomena that grow around its use. The analysis Innis made was not necessarily linear but for the purpose of this essay following these steps will do because my intention is to give a general analysis. Although I am using digital technology as generic term it may refer to different tools and applications. My intention is neither to assess the impacts of specific tools, nor their applications to specific contexts. My intention is to explain how a methodological framework constructed on the notion of bias could function to assess the bias of digital technology and thus come one step closer to understand also the bias of digital preservation. Below I discuss the bias of digital technology and I focus on what could be called cultural bias.

Digital technology for many people refers to the Internet and when applying the Innis method authors usually refer to it because the novelty we experience is triggered by the Internet.²⁸ However, I will not focus on the Internet but rather on the computer because Innis' interest in communication media was his conviction that material factors play an important role.²⁹ However, the difficulty of the analysis lies in the fact that the computer is not just hardware but also software. In any case, instead of describing the roots of the computer, which goes a long way back in time, I would like to explain how the technology reflects the character of the culture that created it and which is evident especially in software. Irrespective of the context in which a medium is created I believe that the context is not just external to the medium but also reflected in it and this context can also be understood by analysing the 'culture in technology'. Each context has its bias and, according to Comor, Innis looked for the reproduction of bias,³⁰ so what bias does the computer reproduce? Answering this question inevitably requires divisions such as Western – non-Western. Although I think that such divisions are overgeneralizations that annihilate the diversity of cultures, in the absence of better terms I will still use them. The best example that can be given to explain the cultural bias of the computer comes from the field of ethno-computing. According to Tedre et al. the history of computer science is an extension of the Western system of knowledge.³¹ "Computers are cultural artefacts that are designed to meet and inherently exhibit the Western understanding of logic, inference, quantification, comparison, representation, measuring, and concepts of time and space, for example."³² One could argue that in recent years computers have developed to be relevant also to non-Western contexts. An example would be the process known in the software industry as "internationalization", which allows software to be adaptable to local conventions, customs, languages, or time zones. However, Mackenzie argues that internationalization software itself is based on erroneous assumptions because, although adaptable in certain regards, "the software itself remains universal in its

²⁷ Cf. Edward Comor, "Harold Innis and 'The Bias of Communication'," in *Information, Communication & Society* (London: Routledge, 2001), 274-294, and Innis *The Bias of Communication*, with Frost, "How Prometheus is Bound."

²⁸ Ibid.; Heyer, *Harold Innis*.

²⁹ Heyer, *Harold Innis*, p. 63.

³⁰ Comor, *Harold Innis*; Innis, *The Bias of Communication*, p. 276.

³¹ Matti Tedre, Erkki Sutinen, Esko Kähkönen, and Piet Kommers, "Is Universal Usability Universal Only to Us?" (paper presented at ACM Conference CUU 2003, Vancouver BC, Canada, November 10-11, 2003), accessed August 20, 2012, http://cs.joensuu.fi/~ethno/articles/ethnocomputing_CUU2003.pdf.

³² Ibid.

aims and expectations because code and software themselves are presumed to be universal as text and as practice.”³³ The underpinnings of software are based on the universality of practices of numbering, enumerating and sorting,³⁴ but as he explains there are differences between cultures, for example between Western numbering practices in base 10, and Yoruba numbering practices, which include base 5, 10 and 20.³⁵ By making reference to research in the field of ethno-mathematics he explains why software should not be considered universal. We can agree that the computer and the logic on which it is based reflect Western culture and thinking; yet this should not deny their potential application and benefits in other contexts. However, Tedre et al. have a point when suggesting that “in searching for something that would be universal or close to universal we need to first understand the local [...] without going to the grass-root level, we will not see the details of cultural influence on technological design, and what we claim to be universals may be just our universals.”³⁶ This is an important step to take to understand the bias of digital technology, especially because, as Innis suggested, the same technology is not the same in different contexts. A technology becomes what it is through its interaction with contextual factors, including cultural factors.

Assuming that the logic of computers is compatible with the logic of other cultures, or rather it would be compatible in the near future if we consider developments in the field of ethno-computing, there are further issues that the Innis method of analysis should include. This leads me to the second step of the analysis, namely assessing the technical characteristics of computers to understand how these influence its adoption and use. Authors that follow the Innis method usually ask questions such as: is the medium easy or difficult to use? Is it efficient in terms of time? Is it financially affordable? Although initially computers were something for the engineer and programmer, they have become a user-friendly technology and thus also a household technology. Following the cultural rationality that I proposed in this essay, and being informed by the previous example regarding the Western logic of computers, I would rather like to ask whether in different cultural contexts computers are still a user-friendly technology. The following example would make us answer this question in the negative. Van der Velden³⁷ explains her experience during a research in the Maasai community. It refers to how a local used certain computer software that was open and flexible so as to be significant for local uses. A local showed van der Velden how he was using it to write articles but he complained that the programme did not include categories that would allow him to meaningfully upload the articles he wrote. Although there was a category for intended audience that included subcategories such as farmers and fishermen, there was no mentioning of Maasai and pastoralist communities, which were his intended audience. Van der Velden explained to him that he himself could create categories that suited him best. However, he refused doing that. He didn’t see it as his responsibility or task because he was not part of the team that created the programme. The conclusion van der Velden draw was that this is an indication of a different understanding of human-technology

³³ Adrian Mackenzie, “Internationalization,” in *Software Studies: A Lexicon*, ed. Mathew Fuller (Cambridge, Massachusetts & London, England: The MIT Press, 2008), 153-160, p. 156.

³⁴ Ibid., p. 157.

³⁵ Ibid., p. 158.

³⁶ Tedre et al., “Is Universal Usability Universal Only to Us?”

³⁷ Maja van der Velden, “Undesigning Culture: A brief reflection on design as ethical practice,” in *Proceedings Cultural Attitudes Towards Communication and Technology*, ed. F. Sudweeks, H. Hrachovec, and C. Ess (Murdoch University, Australia, 2012), 117-123, p. 120.

relations and that it “shows the need for technology designs that allow people to archive their knowledge in a manner that is appropriate to their knowledge and to their way of knowing the world.”³⁸

The third step in an Innisian analysis is to identify the patterns and phenomena that emerge from the adoption and use of digital technology. Innis was interested to see how media involve particular responses, both in terms of social forms, and in terms of cultural psychological predispositions. With regard to social forms, the best examples that could be given are networked and collaborative manifestations such as wikis and blogs. But my interest here is in cultural psychological aspects. Abstract thinking, Innis believed, was enabled by the development of writing.³⁹ So what kind of thinking and cultural patterns does digital technology enable? To answer this question it is useful to turn attention to a new form of text: the hypertext. According to Manovich this type of medium, just like the World Wide Web, is based on the assumption that every object has the same importance as any other, and that everything is, or can be connected to everything else.⁴⁰ He explains that every hypertext reader gets his or her own version of the complete text by selecting a particular path through it.⁴¹ This creates the illusion that people’s choices are not pre-programmed but unique leading to some sort of individualism. Frost agrees that it facilitates individualism and, by making reference to other authors, notes that apart from enabling collaboration and sharing, there are also some concerns that it may lead to the loss of shared meanings.⁴² This hyperlink feature also has cognitive impacts. Nicholas Carr wrote in an article that initially digital documents were believed to have advantages over paper documents: “Hypertext would strengthen critical thinking, the argument went, by enabling students to switch easily between different viewpoints [...] the hyperlink would be a technology of liberation.”⁴³ However, as research developed and started producing results, psychologists showed that the more the links, the lower the comprehension of texts. According to psychological studies this happens because hyperlinks stimulate brain activity in the prefrontal cortex, which is associated with problem solving and decision making, while diminishing the ability for critical thinking and reflection.⁴⁴ This has an impact on how people acquire information and how they understand it. The use of digital technology in the long-term, could thus determine particular predispositions for acquiring knowledge.

Digital preservation, being based on the use of digital technology, certainly takes something from its bias. The analysis provided above, although it is general, gives an impression of where the bias of digital preservation lies and therefore indicates some cultural challenges that should be tackled in order to design more inclusive policies; above all policies that contribute to cultural diversity. So far I avoided intentionally speaking about space and time bias, although this distinction could have been introduced at each of the three levels described above. As I explained before the bias arises from the intersection of various factors and I found it more appropriate to describe these intersections first. Furthermore, the space-time distinction makes sense only if placed in relation with the notion of balance. The chapter below turns now to this aspect but it is preceded by a brief discussion that emphasizes what the bias of digital technology tells us about the bias of digital preservation.

³⁸ Ibid.

³⁹ Innis, *Empire and Communications*, p. 7.

⁴⁰ Lev Manovich, *The Language of New Media* (USA: Massachusetts Institute of Technology, 2001), p. 41.

⁴¹ Ibid.

⁴² Frost, “How Prometheus is Bound.”

⁴³ Nicholas Carr, “The Web Shatters Focus, Rewires Brains,” *Wired Magazine* (June, 2010), http://www.wired.com/magazine/2010/05/ff_nicholas_carr/.

⁴⁴ Ibid.

5. Balance in the Preservation of Digital Heritage

The process of technological obsolescence, which seems to be characteristic of digital technology, shows a predisposition towards discarding the old for the new, or what Innis called “present-mindedness”. This indicates that digital technology is a space-biased medium and most authors see it as such. At the same time they also acknowledge that it has potential as time-biased medium and they motivate this by emphasizing that it enables community and sharing.⁴⁵ It is unquestionable that it has strong leanings towards being space biased; yet its time bias is perhaps discussable because virtual communities are not based on tradition and they are not necessarily durable. Actually I think it is too early to say whether it is space or time biased and my argument is informed by the on-going battle between proprietary and free software. Proprietary software reflects business interests. Private companies have to make sure that they constantly come up with new products and users keep on buying because the business depends on this process. Free software on the other hand is interested in maintaining a community of users not for material gains but for the benefits that free sharing brings. As stated by the author of the GNU Manifesto: “I consider that the Golden Rule requires that if I like a program I must share it with other people who like it. Software sellers want to divide the users and conquer them, making each user agree not to share with others. I refuse to break solidarity with other users in this way.”⁴⁶ This is near to the distinction made by James Carey, who, in a similar spirit with Innis’ space-time bias, distinguished between two types of communication: the transmission view and the ritual view. He says “communication under a transmission view is the extension of messages across geography for the purpose of control ... a ritual view is the sacred ceremony that draws persons together in fellowship and commonality.”⁴⁷ Both aspects are obvious in digital technology but perhaps the mainstream uses people will give it will determine if in the future it will be a space or a time biased medium. Leaving aside the space-time bias, there is a different aspect that I would like to emphasize here: the cultural bias of digital technology that I described above.

In itself the aim of universal access—inseparable from digital preservation—is altruist and informed by equity. But the examples I gave in the previous chapter suggest that digital preservation, despite its generous aims, could have unwanted and perhaps even devastating outcomes for cultural diversity because also digital preservation rests on the assumption that software is universal. Separating the carrier from its content is not a problem as long as we approach preservation as purely technical process. But if we anchor digital preservation in the broader framework of cultural diversity, we cannot but observe that instead of ensuring equal representation of cultures, along time this could lead to standardization of knowledge. Therefore, linking digital preservation with cultural diversity necessarily implies going beyond diversity of contents, and perhaps considering also the diversity of software and hardware. Currently this is not the case in digital preservation. Paradoxically it is standards that ensure interoperability and exchange but at the same time this requires fitting a norm at the expense of the diversity of existing knowledge systems. Perhaps correcting such bias would require, as Tedre et al. suggest, developing universals through particulars, not the other way around.

⁴⁵ Frost, “How Prometheus is Bound”; Heyer, *Harold Innis*.

⁴⁶ Richard Stallman, “The GNU Manifesto,” 1993, accessed August 20, 2012, <http://www.gnu.org/gnu/manifesto.en.html>.

⁴⁷ James Carey, *Communication as Culture: Essays on Media and Society* (New York & London: Routledge, 1989), p. 17.

Cultural bias was also obvious in people's interaction with digital technology. As mentioned above, preservation policies are informed by considerations of user community. For example, the Open Archival Information System Reference Model (OAIS), initially developed by the Consultative Committee for Space Data Systems (CCSDS) but which has developed into a formal standard that has been taken up also by libraries and archives defines a designated community as "an identified group of potential consumers who should be able to understand a particular set of information."⁴⁸ It further states that the designated community may be composed of multiple user communities, which may change over time.⁴⁹ Additionally, OAIS also takes into account that users have a knowledge base that allows them to understand the received information and which should be taken into account by archives. All in all, it is a comprehensive model and it has various advantages that may accommodate to some extent diversity. However, the concept of a knowledge base rests on predefined components and it is informed by the knowledge of the archivist or planner. This may hinder people from interacting with the resource, especially if the users are from non-Western contexts as shown by the example given by van der Velden.⁵⁰ Thus, even if a resource is accessible from a technical point of view, it may not be accessible from a cultural point of view. For interacting with a pre-defined knowledge base, people from non-Western cultures must get acquainted first with a different world-view⁵¹ and this suggests again an erosion of the different worldviews specific to the diversity of the cultures of the world. Considering that the aim of universal access promotes the dissemination of digital technology in different cultural contexts so as to bridge the digital divide, digital preservation, especially in the case of international cooperation projects, would have to dedicate more thought to issues of cultural diversity, if it were to contribute to its protection.

The use of digital technology also triggers specific cultural and psychological predispositions. The Charter raises the problem that attitudinal change has fallen behind technological change,⁵² and this implies that the attitude that has not yet emerged from the use of digital technology is the acknowledgment that digital heritage needs preservation. Using the concept of space and time bias indicates that the problem raised by the Charter comes from the fact that digital technology is currently not perceived as time-biased medium. Furthermore, following the distinction between the transmission and ritual view of communication, we observe that on the one hand the philosophy of digital preservation emphasizes a ritual view of communication. On the other hand its reliance on digital technology indicates that in practice it follows the transmission view. In this regard it is worth mentioning some authors who, although not using the concepts I use, give suggestions that would be suitable to follow the ritual view also in practice. Uricchio for example, talks about new cultural forms that are decentralized, networked, collaborative, and dynamic thus bearing a close resemblance to oral cultures, yet not being like them because they are also embodied in texts and images.⁵³ He suggests that in order to make these cultural forms preservable and accessible archives should follow the logic of these cultural forms, with other words being concerned not only with products but also with the practices.⁵⁴ Owen has a similar suggestion. After describing the

⁴⁸ CCSDS, *Reference Model*.

⁴⁹ Ibid.

⁵⁰ van der Velden, "Undesigning Culture."

⁵¹ Tedre et al., "Is Universal Usability Universal only to Us?"

⁵² UNESCO, *Charter on the Preservation of Digital Heritage*.

⁵³ William Uricchio, "Moving beyond the Artifact: Lessons from Participatory Culture," in *Preserving the Digital Heritage: Principles and Policies*, ed. Yola de Lusenet and Vincent Wintermars (Amsterdam: Netherlands National Commission for UNESCO, 2007), 15-25, pp. 20-21.

⁵⁴ Ibid., p. 24.

characteristics of new digital cultural forms as collaborative and dynamic he says “perhaps the most significant consequence of these characteristics is that modern culture is represented by the use of digital materials and the social and cultural processes they invoke, rather than by the materials themselves. Heritage preservation, therefore, implies not just storage and maintenance of digital artefacts, but the capturing of dynamic processes and patterns of use.”⁵⁵ Doing this would require a reconsideration of basic concepts, including the concept of a document. But perhaps this would not only bring libraries and archives a step closer to knowing how to deal with digital heritage, but could also turn them into institutions that facilitate establishment of digital technology as a time-biased medium.

The analysis provided above indicates that digital preservation will not strive by focusing solely on the technology. There is need to change our concepts and approaches and my particular way of suggesting how such a change could occur was by revealing the cultural bias of digital technology. This would nevertheless be incomplete if we do not also strive for balance, which is the moral message of an Innisian analysis. According to Innis cultures should deal equally with the concepts of space and time, and balance is achieved when two opposing media forms exert parallel influence. In addressing the preservation of digital heritage an Innisian analysis suggests that a balance could hardly be achieved through constant migration of the content. Although pretending to be concerned with time, it indicates an obsession with the concept of space. For Innis constant changes in technology increase the difficulties of recognizing balance let alone achieving it.⁵⁶ The existence of other methods—emulation, encapsulation, technology preservation, and output to non-digital media—indicates that there’s a tacit assumption that entering this game of constant change has its risks. However, an Innisian analysis suggests that each of them has its risks; each is biased in its own way either towards time or towards space. Digital preservation then requires knowing their bias first so as to be able to balance them against each other. We should, however, remember that bias was a complex interaction between different factors not something inherent in technology. Instead of telling something about technology, it tells something about the culture that uses it. If culture is biased towards time, the method of digital preservation would need to tackle the concept of space. If culture is biased towards space, the method of digital preservation would need to tackle the concept of time. There are no universal solutions. The balance can only be found by returning to the human scale and this means understanding culture.

6. Conclusions and Outlook

I have attempted to point out the cultural challenges of digital preservation and, in this regard, I start by relating it to the protection of cultural diversity. I suggest that from a technical point of view the diminished relevance of the carrier may not represent a problem but I argue that this is a problem from a cultural point of view because it draws attention away from the impact digital technology exerts on culture. To support my argument I introduce the theory of Harold Innis. I borrow his concepts of bias and balance and I use them to develop a methodological framework. I use this framework as tool to assess the bias of digital technology and I apply the results of this assessment to the field of digital preservation.

⁵⁵ John Mackenzie Owen, “Preserving the Digital Heritage: Roles and Responsibilities for heritage Repositories,” in *Preserving the Digital Heritage: Principles and Policies*, ed. Yola de Lusenet and Vincent Wintermars (Amsterdam: Netherlands National Commission for UNESCO, 2007), 45-49, p. 48.

⁵⁶ Innis, *The Bias of Communication*, p. 140.

The analysis shows that ignorance of the relation between culture and technology may not only hinder people from using the technology but it may also have negative impacts on culture. This is what I called the cultural bias of digital technology. The analysis further shows that the field of digital preservation, being based on the use of digital technology, has a similar bias and I argue that this is not supportive of the stated aim of digital preservation. Instead of assuring representation of the diversity of the cultures of the world, it may cause its erosion. However, scientific developments such as ethno-computing which place culture at the centre of technology may help correct this bias. Nevertheless, this requires reconsidering basic concepts and approaches. Above all, it requires a balance between different approaches and this balance can only be achieved by understanding the interplay between technology and culture.

Archives Are Not Trees

Hierarchical Representations in Digital Environment

Giovanni Michetti

School of Library, Archival and Information Studies, The University of British Columbia, Canada

Abstract

Preservation means access, and access is key to preservation. Access needs proper description and representation, which will be embedded in the tools used as mediating agents between users and records. XML and markup languages are more and more used to represent archival description in digital environment, so it is fundamental to investigate the hierarchical nature of XML structures and the way it influences the representation of archival materials. The internal structures of archives, articulated in fonds, series and archival units, are more and more represented as trees, and this raises some concerns about the restraints related to the hierarchical model. The adoption of a specific approach in access has consequences on preservation too, since objects' representations stand for the objects themselves, and the name we use to call the thing, will be the thing. Therefore, the complex nature of archives needs to be represented through different and richer strategies.

Author

Giovanni Michetti is Assistant Professor at the University of British Columbia, where he teaches Archival Science. His research area is focused on contemporary archives, and his main research interests are records management, archival description and digital preservation. He has joined several international initiatives on digital preservation. He is deeply involved in standardization processes, as a President of UNI Sub-Committee "Records Management", and Deputy of UNI Technical Committee "Documentation and Information" (UNI is the Italian Standards Organization). He is the Italian representative in a few ISO Working Groups on archives and records management. He authored several articles and essays.

Over the past decade, XML has entered the archival lexicon and found a steady place in the archival tool bag needed to carry out in the digital environment the traditional archival activities, whether they are purely descriptive or primarily managerial.

XML can be seen as a technical means to bypass the limits imposed by operating systems and data architectures, or a concrete and significant contribution towards interoperability, or a strategic resource for long-term preservation, or a spur to reflection on the definition, articulation and granularity of the information elements supporting archival systems: XML has undoubtedly many different and interesting profiles. To correctly interpret and adequately address the evolutionary processes taking place in our domain, we must understand the role of this technology and its potential. In particular, I'll focus on the hierarchical nature of XML constructs, so evident that makes it quite natural to use the term *tree* to refer to the articulation of *elements*, the building blocks of XML, and hence—by abstraction—to the internal articulation of documents and their aggregations.¹

¹ Actually, on the strength of the deep connection, rather, the matching we progressively establish between representation language and represented reality, we may legitimately adopt the opposite interpretation in which trees are seen as abstractions of documents or document aggregations, expressed through XML. The point at issue here—the doubt that drives our discourse—is whether this matching is adequate.

The tree as an effective iconic representation; a tool for *viewing* a document or an archive. The tree as a seemingly obvious metaphor, related straight to the daily experience of folders and sub-folders by which we organize the digital documents on our computers, an experience which in turn may be traced back to the logical and physical architecture which governs archives and libraries from ancient times. The tree as structure; in other words, as a representational model of the complex reality of a body of documents.

Nevertheless, the superficial simplicity of these images cracks immediately under the weight of terminology and concepts of dense meaning. In fact, if we assume that *structure* is, in its most generic meaning, “a whole, the parts of the whole and the relationship of these parts to each other,”² or—in other words—“a system in which everything is connected,”³ always with the understanding that structure is *both* the whole connected structure and the system of connections; assumed this definition, can we be satisfied with a model based upon hierarchical relationships only? Can we state that the tree properly represents the portion of the world that is the object of our interest? In short, shunning the awe for our masters and reinterpreting their lesson: is it really true that trees mirror archives?⁴

Claudio Pavone writes in his famous article “arranging an archive means placing its individual components in meaningful, mutually related positions. The meaning comes out [...] from the order itself — in other words, it is connected with the formal structure of the archive and made explicit by the inventory, rather than with the content of the individual components.”⁵ And there is no mention of the fact that these relationships must be hierarchical; rather, the only constraint is that they have a meaning indeed, thus admitting the possibility of establishing connections otherwise.

Where from, then, this tendency to think *vertically*?

The question might seem specious— if not irreverent—because it questions the paradigm *fonds - series - archival unit* that—expanded, reduced or otherwise elaborated—is the common way of thinking about the archive, the model used in thousands of inventories and finding aids. However, the point that should be stressed—and too often slips out—is that an inventory contains knowledge concerning a certain reality, and we deliberately use the generic term *reality* to denote both the whole archival body and the context—be it historical, institutional, cultural, social, administrative or any other kind—in which it is immersed. In traditional (paper) finding aids such knowledge is transmitted through linguistic-rhetorical or—broadly speaking—semiotic techniques and stratagems, through which we can represent the network—better: the nebula—of concepts, events, places and people that contribute to define the environment in which the described *corpora* are set.

Therefore it is true that what we have metaphorically called *vertical trend* belongs to the experience and methodology of our discipline, but we have to recognize that it has so far been mediated—perhaps we

² My translation. Umberto Eco, *La struttura assente. La ricerca semiotica e il metodo strutturale* (Milano: Bompiani, 1998), 256-257. Original citation: “un insieme, le parti di questo insieme e i rapporti di queste parti.”

³ My translation. Eco, *La struttura assente*, 257. Original citation: “un sistema in cui tutto è connesso.”

⁴ The Author refers here to a fundamental article written in 1970 by Claudio Pavone, a master of archival science in Italy, where he argues about a key principle of the Italian doctrine, i.e., the match between archives and creators. The title of the article may be roughly translated as “Is it really true that archives reflect organizations?” See Claudio Pavone, “Ma è poi tanto pacifico che l’archivio rispecchi l’istituto?” *Rassegna degli Archivi di Stato* 30, no. 1 (1970): 145-149.

⁵ My translation. Pavone, *Ma è poi tanto pacifico*, 148. Original citation: “Ordinare un archivio significa collocarne i singoli pezzi in posizioni reciproche e collegate che abbiano un significato. La significatività scaturisce [...] dall’ordine stesso; è cioè connessa alla struttura formale dell’archivio, resa esplicita dall’inventario, e non al contenuto documentario dei singoli pezzi.”

should say constrained—by a presentation that has always brought back the heritage of knowledge about an archival entity into the scope of a narrative hopefully able to represent it as a *nebula*. What's more, in traditional finding aids even superficial understanding of an archive is strongly influenced by the reading of the finding aid, and such a reading—at least in the first instance—occurs mostly following the sequential patterns learnt on printed text.

In digital environment, the fragmentation of information—which offers many advantages in data research and recovery—has progressively moved away those mediating factors that help to *read* an archive as a *story*—rather, a novel; even better, a mystery—where every element, every paragraph, every detail, to varying degrees, is useful and necessary to find the murderer.⁶

The fragmentation of information has reduced the narrative in favor of exasperation of structure, causing the illusory feeling that mastering the tree is equivalent to mastering the archive. Also, this fragmentation has suggested new possibilities of *reading* the archive, released from the sequential approach, thereby facilitating the creation of new and more elaborated research paths, but introducing the obvious risk inherent in all analytical processes that are not followed by an appropriate and decisive process of synthesis: the inability to bring the many components to an organic unity, the individual details to the whole figure that defines the overall appearance of the archive.

Thus, the transition to digital environment—where not even reduced to trivial electronic photocopying of traditional finding aids—has caused a decline in the use of rhetorical and linguistic techniques in favor of a more schematic approach, a reduction of the narrative and a simplification of the network of relationships, all phenomena induced by the use of certain technologies and data structures.

That's the real root of this *tendency to vertical*. The digital environment has been the breeding ground of this drive, then exasperated by markup languages. XML is inherently hierarchical, intrinsically designed to create trees, rigidly constrained by the classificatory logic of inclusion.

Therefore, the image of a tree is much more than a vague suggestion, it rather refers—even from an iconic point of view—to the *Porphyrian tree* that has been object of philosophical debate from antiquity to the present day:⁷ a scheme that interprets and translates graphically the doctrine of the Aristotelian categories, a logical system of subordination where from supreme genus you descend down to lowly species, and hence to individuals⁸ (what we now call *instances*, borrowing the term from computer science). The Porphyrian tree is a model of knowledge representation, based on the fundamental concept of *specific difference*, that is, the identification of a character acting as a crucial element to uniquely determine the species. In fact Porphyry distinguishes between *constituting differences*—those that define

⁶ Actually, the murderer is already known: it's the archive itself, as Jacques Derrida writes (with a psychoanalytic connotation). "Right on what permits and conditions archivization, we will never find anything other than what exposes to destruction, [...] introducing, a priori, forgetfulness and the archiviolithic into the heart of the monument. [...] The archive always works, and a priori, against itself." Jacques Derrida, "Archive Fever: A Freudian Impression," transl. Eric Prenowitz, *Diacritics* 25, no. 2 (Summer, 1995): 14.

⁷ References to the Porphyrian tree are disseminated through scientific literature on bibliographic classification, while studies in the archival domain nearly don't mention it. We are aware of the gap between disciplines and activities that are related but not identical. Nonetheless, we feel the lack and the need for a purely theoretical reflection on the concept of archival classification, able to suggest a consistent theory connected to the fundamental acquisitions of the philosophical, bibliographic and—last but not least—semiotic domain.

⁸ In the example provided by Porphyry himself: from substance to body, hence animated body, and then man, the lowest species before we find individuals ("Substance indeed, is itself genus, under this is body, under body animated body, under which is animal, under animal rational animal, under which is man, under man Socrates, Plato, and men particularly."). See Porphyry, *Introduction (or Isagoge) to the logical Categories of Aristotle Isagoge*, trans. Octavius Freire Owen (1853), 614.

the species—and *splitting differences*, as those that trigger the mechanism of subordination, i.e., filiation within genera and species.

It is easy to recognize in this model the philosophical foundation of well-known taxonomies, such as that of Linnaeus, who moved the hierarchical classification from the logical-ontological to the biological domain; or those more familiar to us, such as bibliographic classifications (e.g., the Dewey Decimal Classification).

But above all, by the analogy proposed it is easy to recognize that the adoption of XML is more than just a technical option: it is rather the choice of a specific knowledge paradigm, not at all neutral—as indeed true for any technology—but rather intrinsically acting as an information tool, a stratagem to give form and model reality. In a sense, we could say that markup languages are before and beyond the real world objects they are meant to describe, so they may be thought as tools for interpreting the reality. But at the same time, just as languages, they are a representation of the world and draw from it the design of their syntactic structures.

This apparent contradiction is actually the dynamic factor that drives the cognitive process by which we aim at identifying the nature of archival objects and their relationships; it is a hermeneutic process that it would be simplistic to simply define as hypothesis and testing—these move just from the real they intend to analyse—and seems best described as a permanent tension between postulation and discovery: we do not just find and discover structures in things—often is not even possible a precise examination of all objects being analysed, because of their variety—we rather pose the structure *ex-ante*, we create it as hypothesis and theoretical model, postulating that the phenomena under study correspond to the theorized structural arrangement.⁹

Continuing to think in pictures: to work in an XML environment forces to some extent to leave a pledge, that is, to compel the descriptions to the expressive capabilities offered by XML. And although the hierarchical logic seems the ideal way to narrate the structures that inform the documents, nonetheless it fails to fully represent—even at document level, i.e., the atomic content of a corpus—the system of relations that binds the elements in which the object is articulated. In fact, it is true that DTDs and XML Schema can provide accurate formal descriptions of a document structure: using a formal language they strictly define the constituent elements of the tree, their internal articulations and their interdependencies. Nevertheless the problem lies in the definition itself of the concept of structure: such an XML tree tells us nothing—except for what we get by intuition—about the meaning underlying the subordination of one element to another.

Let's consider for example an element `<document>` articulated in `<protocol>`, `<text>` and `<excatocol>` (Figure 1); and an element `<title>` articulated in `<original_title>` and `<given_title>` (Figure 2).

It is clear that in the first example the children of `<document>`¹⁰ do not belong to the class identified by `<document>`, but are the actual components of the `<document>` element: a `<protocol>` is not a subset of `<document>`, and the same is true of `<text>` and `<excatocol>`. The formal aspect should not be confused with the substantial: they may well exist written objects without—let's say—protocol or excatocol, because they are damaged or written according to a very different compositional logic, and yet these are to be considered documents to all effects. However, in reference to the archetype, to the *form* of document—which in a sense pre-exists to the specific documents that may be found in specific

⁹ These words echo on purpose Eco's words about structure. See Eco, *La struttura assente*, 49.

¹⁰ For the sake of precision, we should refer to the *classes identified by the elements that are children of <document>*. Hereafter, we will adopt a shortened form, when needed to improve readability.

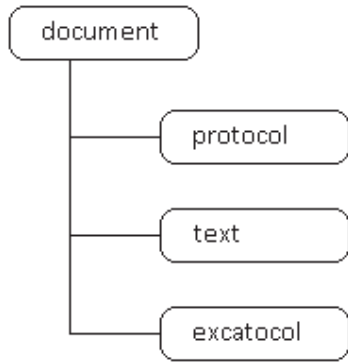


Figure 1.

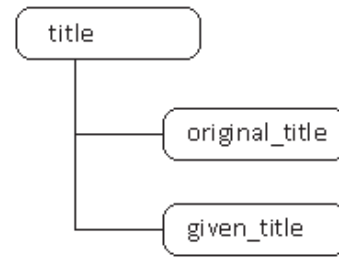


Figure 2.

archives—it is not possible to say that a <protocol> is a <document>, in the same way in which we could not say that a <date> is a <document>.

While in the second case, both <original_title> and <given_title> may well qualify as titles, i.e., they are specializations of the genus <title>. In both cases, we are dealing with *abstractions*: the former is a so-called *aggregation*, the latter, a *generalization*.¹¹ They are different but there is no difference in the tree: a node with three children, in the first case, a node with two children, in the second. Nothing else.

To put it another way, the semantics of the relationship is lacking, and remains suspended in the air, left to the reader's interpretation.

A further element of consideration: the XML tree establishes hierarchical relationships that identify dependencies that are either *immediate* (or rather, *un-mediated*, i.e., related to a direct relationship between a node and its children) or *mediated* (i.e., related to the relationship between a node and an ancestor that is not in the straight line of succession in the genealogical tree). How to describe the relationship between two nodes that belong to disjoint branches of the tree? That is, two nodes that can only be connected by walking the tree along a path that includes ascending and descending stages? The XML tree does not allow horizontal or oblique navigation (and relationships); it only allows strictly vertical navigation. This makes it difficult to represent certain meaningful connections in an *immediate* way, and ultimately makes it impossible to represent a portion of knowledge about the objects.

Developing the previous example, let's suppose we have identified the element <chronological_date> within the <protocol>, and the element <topical_date> within the <excatacol>, between the end of the text and the subscriptions,¹² as illustrated in Figure 3.

The <chronological_date> and the <topical_date> would be seemingly unrelated nodes, placed indeed at the same depth (both children, the former of <protocol>, the latter of <excatacol>, and both grandchildren of the same root element <document>), but separated by a logic gap in evident

¹¹ *Generalization* refers to an abstraction in which entities of different types are considered as examples of a higher level entity. *Aggregation* refers to an abstraction in which entities of different types are related through a relationship defining a higher level entity.

¹² Such a model is not outré: the *datatio* (see next note) is usually placed after the *excatacol* in public documents, while “in private documents is usually either placed in the protocol, after the *invocatio*, or split in two, so that the chronological date is placed in the protocol and the topical date at the beginning of the *excatacol*,” Alessandro Pratesi, *Genesi e forme del documento medievale* (Roma: Jouvence, 1987), 87-88 (my translation). Original citation: “nei documenti privati trova posto di solito nel protocollo, dopo l’invocazione, oppure viene divisa in due membri, di modo che la data cronica risulta espressa nel protocollo e quella topica all’inizio dell’escatocollo.”

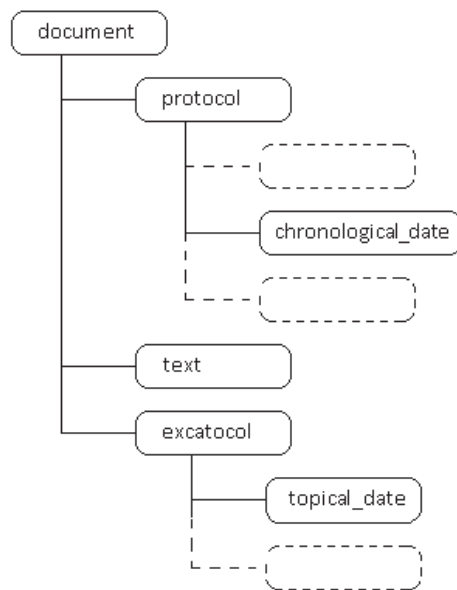


Figure. 3

contradiction to the unity of their function and meaning. This gap can be reduced and eliminated only through the user's text culture, which will allow them to interpret correctly the aseptic hierarchical representation of the structure, that is, to bring back to unity (the *datatio* as a whole¹³) what the tree represents as separate.

Moreover, the XML tree identifies a *structure*, as we said, an articulation of the elements that make up the skeleton of the object of analysis, with the ambition to represent nothing more than a system of relations. In other words, the XML tree should refer to syntax only, to such an extent that we may consider as equivalent—for example—the models in Figures 4 and 5.

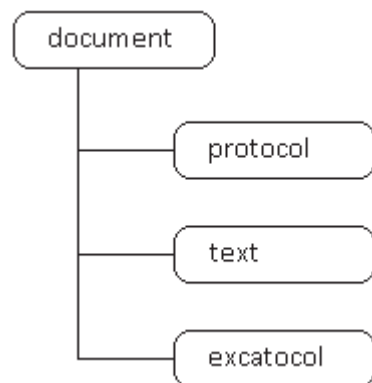


Figure 4.

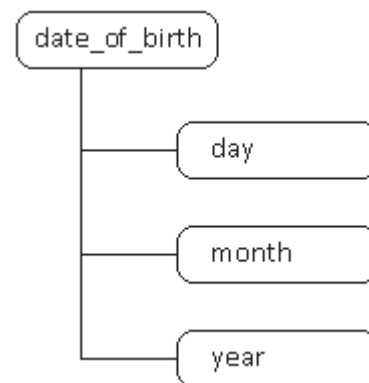


Figure 5.

¹³ According to Diplomatics, *datatio* identifies the statement about the date. It includes information about both time and place of writing.

There is in fact an *isomorphism*, i.e., a transformation which brings the two structures to overlap each other and coincide, highlighting the form (it almost comes to say *Aristotelian* form) of the model, which ultimately may be easily illustrated through a statement that in natural language would *narrate* the model as “three different elements directly subordinated to the same element” or “element X has elements Y, W, Z directly subordinated.” But it is obvious that such an abstraction would destroy in a single stroke all the implicit knowledge nested—willingly or not—in the names of the elements: because it’s all very well if we state that XML is just *syntax*, but the truth of it is that we intuitively associate to the <protocol> tag exactly the notion of *protocol* that we have learnt studying Diplomats and reading ancient parchments; as well as we expect the <year> element to convey chronological, not—let’s say—topological information.

The choice of self-explanatory names precisely aims to highlight the nature of an element, so as to make it *recognizable* without the need for definitions. But this is *semantics*: doing so we surreptitiously attach *meaning* to the model, hiding it within the folds of the labels associated with the nodes. It could not be otherwise: to completely isolate structure from semantics seems an operation doomed to failure, as we have seen. We need semantics to denote relationships, even if they are only hierarchical, and we need semantics to denote the elements of a structure. Note—incidentally—that in the example above we use the term *protocol* to mean a specific structural element that is included in the introductory part of a document. The same term can also be understood—still within the archival domain—in a different sense, to identify “a diplomatic document, especially a treaty or compact, signed by the negotiators but subject to ratification.”¹⁴ What allows us to understand what is the proper meaning? Certainly not the structure, but rather the context of use: again here we are, discovering traces of meaning into structures.

Things are not obviously better when we move from a single document to a collection: not only the problems are replicated on a larger scale, but are also made more complex by the multiplicity of relationships, both within the confines of a single body and—more broadly—within a *corpus* that is intrinsically made of those relationships, and should be interpreted and modeled as a forest of trees, a cohesive entity that does not deserve to be reduced to a sum of trunks.

Further problems arise when you want to represent relationships between entities that are not homogeneous: so far we have considered just documents or—at most—aggregations of documents, and we have recognized some notable limitations imposed by hierarchical relationships. A fortiori, it is clear that the relationship between creator, curator and archives—just to mention the most important entities that make up the archival universe—is not generally ascribable to a model of hierarchical constraints.

To finally complicate the picture, it should be noted that the form itself of a system of relationships actually communicates meaning: let’s consider for example a filing plan regardless of the names of the individual items; in other words, let’s consider its skeleton, what above we have called *form*. Well, the very articulation of the items provides knowledge: a very *deep* filing plan, i.e., extended much more vertically than horizontally, with a lot of first-level items and a few lower-level items, expresses, or rather, represents:

- The variety and richness of the activities and functions of an organization
- The unfitness if not impossibility to aggregate those activities and functions in broader categories
- The limited range of processes related to each activity and function.¹⁵

¹⁴ Richard Pearce-Moses, *Glossary* (Chicago: Society of American Archivists, 2005).
<http://www2.archivists.org/glossary/terms/p/protocol>.

¹⁵ In a similar but opposite way, we may interpret a very *wide* filing plan, i.e., extended much more horizontally than vertically, with many different levels each with very few items.

A good amount of relevant information indeed. Knowing nothing of the “content” of a filing plan, but relying only on the structure, which provides basic information, in a way decisive to correctly interpret the data that populate the system in which the filing plan is supposed to operate.

To paraphrase McLuhan we may dare a hyperbole and state “the structure is the message.” And as in all the hyperboles, we would get a grain of truth: a dictionary, a phone book, a protocol register, a library catalog—all strongly-structured objects—communicate immediately a message, if not a function or usage. In a sense, their structures determine the content. In fact—and this is the heart of the problem—the structure itself is content, becomes content, determines the content. To put it in more fascinating words, “the archive is not only the place for stocking and for conserving an archivable content *of the past* which would exist in any case, such as, without the archive, one still believes it was or will have been. No, the technical structure of the *archiving* archive also determines the structure of the *archivable* content even in its very coming into existence and in its relationship to the future.”¹⁶

In other words, rules, descriptive models, data structures affect decisively the very *identity* of the objects we archive, because circumscribe and limit the *space* of possibilities to create and represent the system of relationships in which those objects are put, the same space that—to a careful examination—we recognize as defining the identity of the objects. Therefore, the question at the origin of our thinking takes a more challenging nuance, since it can be related not only to description and representation of an existing and well-established archive, but also to its creation, and to the creation of its identity. So we should not only ask if the tree mirrors the archive, but even if the tree is the appropriate structure to *create* the identity of the archive, being aware that “what is no longer archived in the same way is no longer lived in the same way. Archivable meaning is also and in advance codetermined by the structure that archives.”¹⁷

In summary, the considerations carried out so far have highlighted the criticality of the hierarchical model, showing that a tree may represent only part of the knowledge embedded in or related to an archive.

So why not get rid of the *prejudice* that recognizes hierarchy as the privileged relationship? Why not build instead a predicative model where entities are loosely connected and connectable to each other through relationships having a different nature? Why not explain the semantics we have seen hidden in the folds of the tree, for example using definitional and classificatory schemes formally identified and defined, and unambiguously associated with the entities in the model? In short, why not build an *ontology*?¹⁸ Why not design a system where the entities we *recognize* as having the right to exist, and the allowed relationships are defined with formal rigor?

Consider also that the natural environment for the implementation of the considerations carried out so far is not usually the limited space made up by one or more computers, or an isolated and self-sufficient information system. Nowadays the Internet is the medium usually adopted to disseminate our digital representations. The tree then, but on the Web, so it appears more vivid the contrast between the simplicity—not to say triviality—of the hierarchical structure, and the complexity of the network model. Why delegate to a vertical design what is then immersed and communicated through a medium which invades the space of our action with the complexity of a neuronal network?

¹⁶ Derrida, “Archive Fever,” p. 17.

¹⁷ Ibid., p. 18.

¹⁸ In information science, an ontology is an explicit formal description of concepts and related properties (including relationships) within a domain. An ontology allows one to model a specific portion of reality (the domain) using not only subsumptive (hierarchical) relationships, but also more complex semantic relationships through which the connections among concepts can be described accurately.

Even from the mathematical point of view, it is worth noting that a network is a graph: nodes connected with other nodes. And a tree is a special type of graph. Why reduce the degree of complexity of the representation structures? Why choose simpler structures if they are then immersed in an environment that can support complexity?

If we accept the fact that technologies not only change the way we communicate, but also transform the public and private spaces of our societies, redraw their borders, impose changes of legal and political nature, in short *shape* our culture and ideologies, why not lead this paradigm to the end and adopt for our discipline structures, tools and strategies consistent with this approach?

XML is neither *the* problem nor *a* problem. Rather, it is part of the solution. But it is only a starting point. It is an environment, a territory within which we should work, a toolbox to create information architectures consistent with our view of reality. The world as we represent it, is an effect of interpretation and our reflections here aim to put in discussion not the act of interpretation itself, since it is a hermeneutical ineludible activity, but the expressing structures that are supposed to represent and to mirror that interpretation.

The tree therefore as a starting point, as an instrument to (partially) represent an archive, as a grid to sift through the treasure hidden in the archives. Being aware that by varying the grain and texture of the sieve we can discover and retrieve parts of that treasure otherwise escaped from our control.

The New Information Landscape

The Archivist and Architect – Drawing on a Common Map?

Göran Samuelsson

Mid-Sweden University, CEDIF - Centre for Digital Management

Abstract

For a long time the archival profession has argued with great tenacity for the importance of working proactively. Nevertheless, we are still far from achieving this goal. In today's environment in which an increasing volume of digital information is used by more and more over longer periods of time, the need for orderliness is increased. How can we get a better map of today's digital information flows and simultaneously assure the quality of information over time? This paper highlights the need to focus on an overall information management from the first sketch of an information system to eternity. To create the conditions for a good information management it is necessary for archivists to improve their knowledge in information architecture and establish liaison with enterprise information architects. This paper aims to develop a mutual understanding for both disciplines' techniques and methods and suggests ways they can work together to develop even more efficient information management strategies and methods.

Author

Dr. Göran Samuelsson is Assistant Professor in Archives and Information Science at Mid Sweden University and Project Leader for CEDIF, the Centre for Digital Information Management. Earlier he worked as an archive coordinator at the National Land Survey of Sweden, dealing with organizational, strategic questions and large digitization projects. His research interests include storage and long-term preservation of records in the digital environment; recordkeeping systems dealing with geospatial information; and education and professional development for the archives and records management community. He is a member of the Swedish ISO/SIS group for records management, the Swedish representative for European Spatial Data Research (EuroSDR) and a member of the Data Archiving Working Group.

1. Introduction

For a long time the archival profession has argued with great tenacity for the importance of working proactively. Nevertheless, we are still far from achieving this goal. This article aims to highlight the problem and will suggest some ways to achieve these objectives sooner. It draws from research undertaken within the Centre for Digital Information Management (CEDIF) at Mid Sweden University.¹ The project had four main tasks. To:

- Develop and establish a centre for archives and information management at the Mid Sweden University
- Develop models for electronic information management (Research part)
- Develop a digital systems laboratory
- Disseminate and communicate experiences, knowledge and results.

¹ www.miun.se/cedif

This article will primarily focus on phase (b) dealing with research. The research part of the project had the ambition to cover the total range of information management in seven different aspects (see below) whereas the main focus for this text is on the first dot points:

1. Enterprise content management & enterprise architecture
2. Business process management
3. Documentation
4. Records management/ recordkeeping
5. Metadata
6. The borders between records management and archives management: The archive as a function
7. Systems for long-term preservation

In one way this structure with seven dots also reflected the information flow and its way from a think/planning phase to management of information for eternity. This ambition emerges among others from the strong growth in digital information. In the last decade we have seen an increase and widespread re-use of information through different e-services. Another drivers had been legal aspect in the finance sector as SOX² and the European Basel,³ collecting and gathering of information in the area of Environment and Health. Even the European PSI-directive (Public Sector Information 1 July 2005)⁴ has actively influenced the business to create more structured and complete digital information flows for access. Then we have the more internal and business-driven reason for the growing amount of information. The Business Applications has striving towards fully digital management and an online e-services 24 hours seven days a week. The need for the business to support long-term customer and business relations create information to be kept for many years. Using older data for statistics and analysis is a growing business either it's called Business Intelligence or Big Data. The storage industry expects the information to grow 50 times compared with today to year 2020.⁵ All this information will require more storage space. The only thing we with some certainty can say about the future of everyday life is; an increasing amount of information that is used longer by growing numbers of people. Therefore it is becoming increasingly important to have good order at an early stage. To come from a stage where the business process looks more like a mess (Figure 1). To a more analysed and structure way is the goal for many business today (Figure 2).

2. The Archivist's Mission - tools and methods

Before I describe in more detail the (Swedish) archivist's method and approach to creating this "good order at an early stage" you need some knowledge about the settings in a Swedish context and about the Swedish framework. First and quite important, why create and keep archives? In Sweden we have the Archive Act,⁶ stating that public records must be preserved, kept and managed so that they meet the rights of:

² Sarbanes–Oxley, Sarbox or SOX, is a United States federal law that set new or enhanced standards for all U.S. public company boards, management and public accounting firms. It is named after sponsors U.S. Senator Paul Sarbanes (D-MD) and U.S. Representative Michael G. Oxley (R-OH).

³ Basel III is a global regulatory standard on bank capital adequacy, stress testing and market liquidity risk agreed upon by the members of the Basel Committee on Banking Supervision in 2010-11.

⁴ http://ec.europa.eu/information_society/policy/psi/actions_eu/policy_actions/index_en.htm

⁵ http://chucksblog.emc.com/chucks_blog/2011/06/2011-idc-digital-universe-study-big-data-is-here-now-what.html

⁶ The Swedish Archive Act (SFS 1990:782 §3)

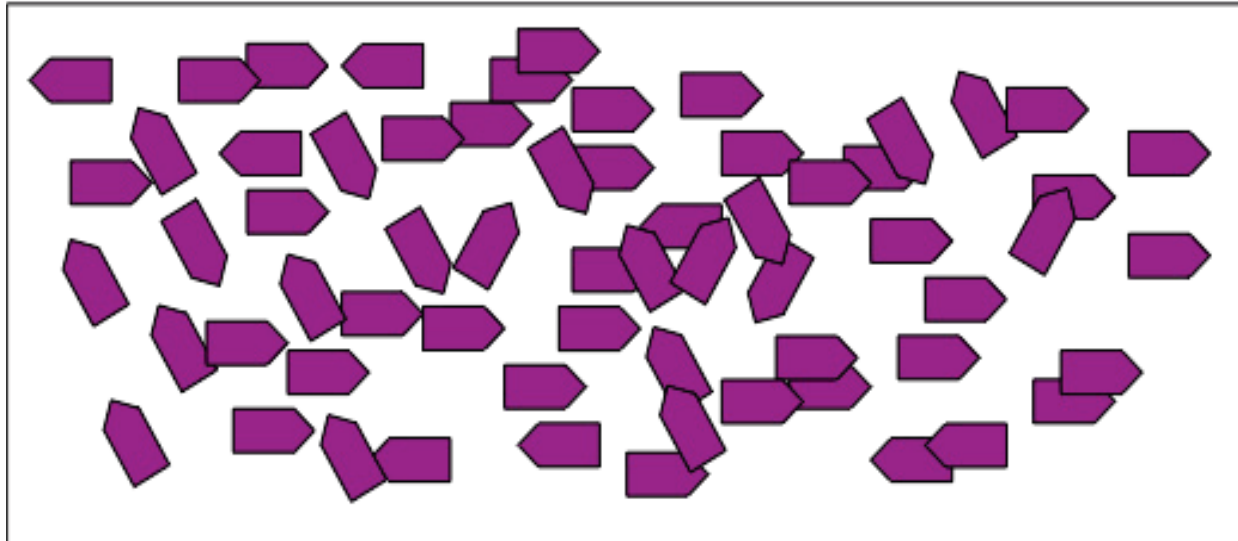


Figure 1. Information flow mess.

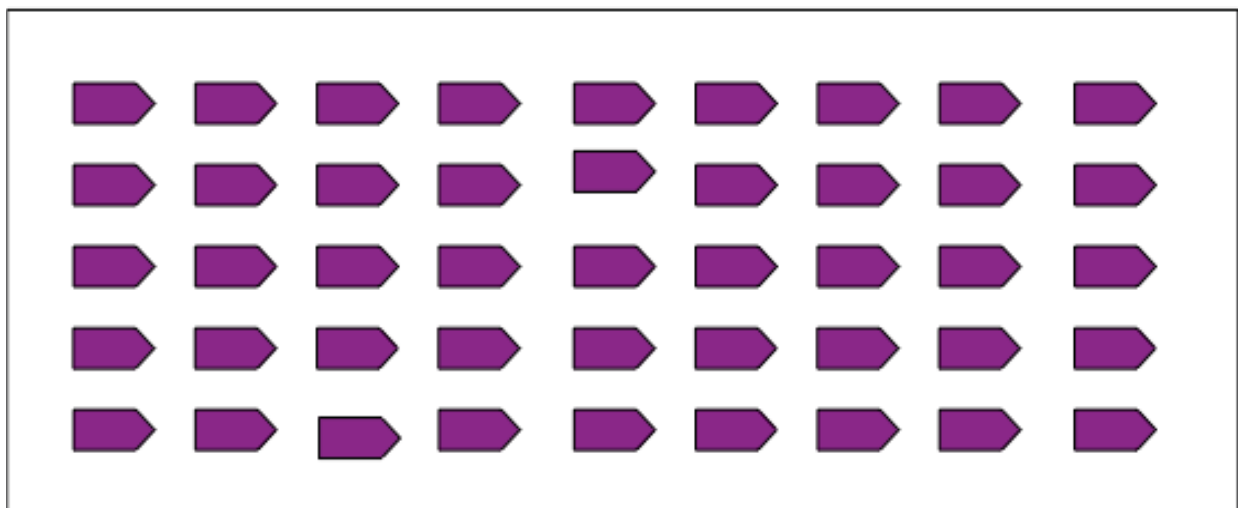


Figure 2. Information flow good order.

1. Access to public records,
2. The need for justice and administration
3. Research

It indicates strong support for transparency and control of government functions. Since 1766 Sweden has had a Freedom of the Press Act. It provides each and every one of the Swedes free access to all records of central and local government. This is called the Principle of Public Access (Offentlighetsprincipen). And in an archival context it might also be important to add that Sweden doesn't distinguish between Records managers and Archivists. We have just one profession dealing with the information from creation to eternity—the archivist.

In 2009 the Swedish National Archives published a regulation for dealing with electronic information and records.⁷ The regulation stresses that the electronic information should be useable for eternity. “The electronic records shall continuously be produced, transmitted, recorded, handled, stored and managed so that they can be presented repeatedly during the time that they will be preserved...”⁸ These activities shall be executed in accordance with the strategy stated in chapter 3 “Strategy and planning for preservation of electronic documents.” The first and second paragraph in that chapter say that an “Agency shall establish a strategy for preservation of electronic documents... which should be continuously supplemented and updated.”⁹ The documentation must be presented in a cohesive way and also linked to or integrated with the agencies’ archival descriptions.¹⁰ This archival description has its own regulation (RAFS 2008:4), which should make it possible to:

- Understand the relationships between business and actions,
- Give a total overview of the documentation
- Make it possible to search in this documentation
- Support the handling and management of these documents.

The archival description must cover all the Agency’s documents and also have a classification structure with process descriptions and an archival inventory.¹¹ In the next part we will have a closer look at the method and tools to master these requirements.

3. Methodology for business analysis and document classification

Swedish archivists are now required to use a process-oriented description system based on their knowledge of how business information is generated within the organisation. The basis for this work has often been to use the information and documentation plan for the actual business. The documentation plan gives authority for the necessary controls and oversight of the public documents, facilitates the archivist’s work and improves the public’s ability to access and re-use information. The plan indicates what processes and activities exist and what documents are created and managed in them; records the document types, formats, media and appropriate storage strategies for them; and which records are protected by confidentiality, retention periods, etc. I will shortly describe the different steps in the creation of this plan. These steps closely follow the first part of ISO 15489 (Step A: Preliminary investigation, Step B: Analysis of business activity, Step C: Identification of recordkeeping requirements). First, the business purpose and mission, etc., is documented. Then the most important processes are documented, activities are described and an overarching process map is created, preferably also documenting the processes graphically (Figure 3).

3.1 Requirements for records

After these first steps it’s time to focus on information about the core business processes and connect the records/information objects to functions, processes, activities and transactions (Figure 4).

⁷ RA-FS 2009:1 (Swedish National Archives regulations)

⁸ RA-FS 2009:1 Chapter 4 - 1 §

⁹ RA-FS 2009:1 Chapter 3 - 2-3 §

¹⁰ RA-FS 2009:1 Chapter 5 Documentation 1 §

¹¹ RA-FS 2008:4

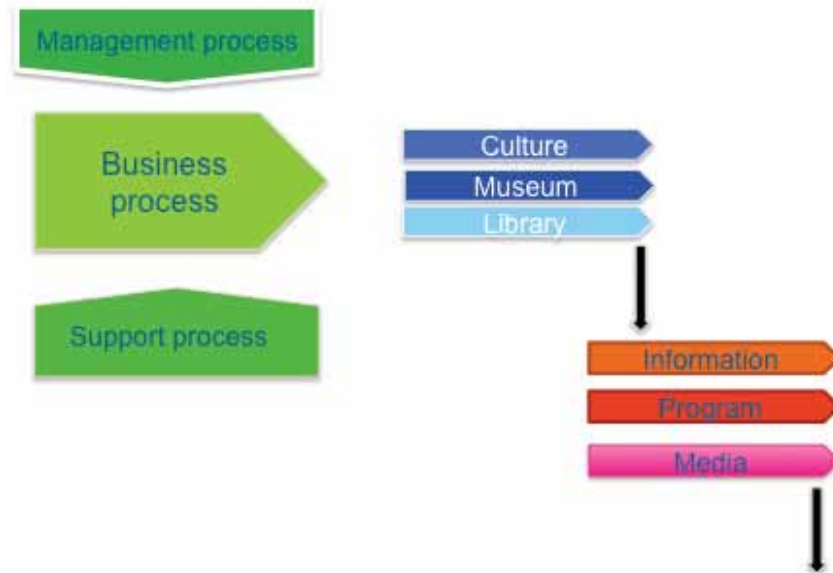


Figure 3. Overarching process map.

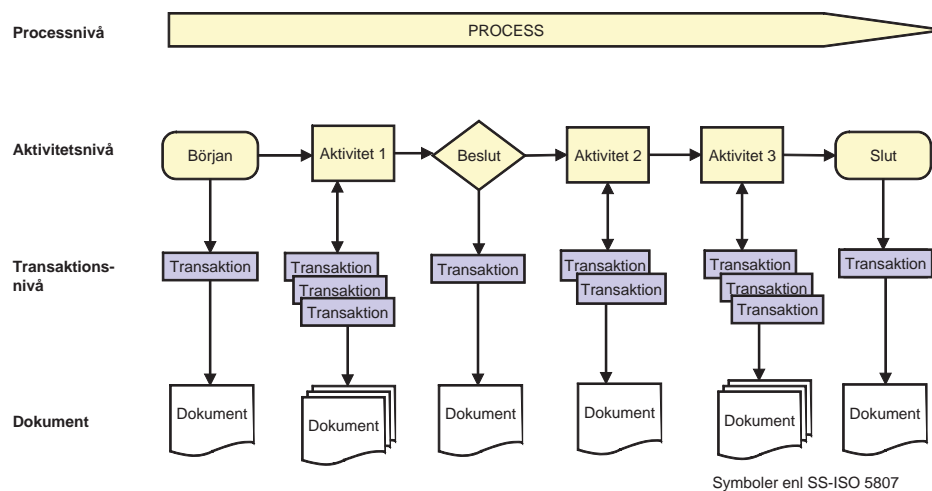


Figure 4. Processes and information objects.

At this point the archival format and storage medium for digital documents is decided, and appraisal and disposal is determined. The criteria for sufficient metadata must also be defined. This results in a document-plan or information-plan where all essential data about the business information assets will be documented (Figure 5).

4. Architect - Mission, tools and working methods

Parallel to these “archival” developments, a new cadre of information workers has emerged over the past two decades. An overarching description for them could be ‘IT-architect’ and although the ‘T’ in ‘IT’ was initially over-emphasised by this new group, the importance of information has increasingly come into

Nämnd	Kultur- och fritidsnämnden							
Verksamhet	Kultur - Museum - Bibliotek							
Processområden	Information - Program - Mediaförsörjning							
Processer/Aktiviteter:	Inköp av media - Katalogisering - Libra-registrering - Cirkulation - Fjärrlån - Mediagallring							
IT-stöd	Officepaketet, Libra							
Senast ändrad					Dokumentombud Bengtsson			
Process (rubrik)	Registrering D/P/V/S	Analoga dokument		Digitala dokument		Bevaras eller Gallr- frist	Leverans- frist till Kommun- arkivet	Format vid arkivering
-Handlingsslag -Handlingsslag -		Format	Förvar	Format	Förvar			
Inköp av media								
Sambindningslistor	S	P	Avd			1 år	-	
Inköpsförslag och gallringsförslag	S	P	Avd	(Word)	(N)	Vid inakt	-	
Minnesanteckningar, tidskrifts- inköpsmöten	S	P	Avd	Word	N	10 år	-	
o s v								

Figure 5. Document Plan.

focus. This group originated as an information technology (IT) discipline. Initially it focused on promoting the strategic development of an organisation's IT systems by modelling the organisation and by aligning IT purchasing and development with business priorities. In recent years the scope of enterprise architecture has expanded beyond the IT domain and enterprise architects are increasingly taking on broader roles relating to organisational strategy and change management.

It was around the year 1987 when this field that soon came to be known as enterprise architecture (EA) was born. The field initially began to address two problems:

- System complexity - organisations were spending more and more money building IT systems; and;
- Poor business alignment - organisations were finding it more and more difficult to keep those increasingly expensive IT systems aligned with business need. An IT system was unmanageably complex and increasingly costly to maintain. They hindered the organisation's ability to respond to current and future market conditions in a timely and cost-effective manner. Mission-critical information consistently was out-of-date and/or just plain wrong.

Many times almost a culture of distrust emerges between the business and technology sides of the organisation (Figure 6).¹²

4.1 Some definitions of EA

Regardless of perspective in different EA-traditions, the idea is to establish a more holistic perspective over the organization with a mission to facilitate more efficient management.¹³ The Gartner group had

¹² Roger Sessions, "A Comparison of the Top Four Enterprise-Architecture Methodologies," (2007), <http://msdn.microsoft.com/en-us/library/bb466232.aspx>.

¹³ T. Tamm, P. B. Seddon, G. Shanks, and P. Reynolds, "Delivering Business Value Through Enterprise Architecture," (2011), <http://toomastamm.com/cv/files/BusinessValueOfEA.pdf>.

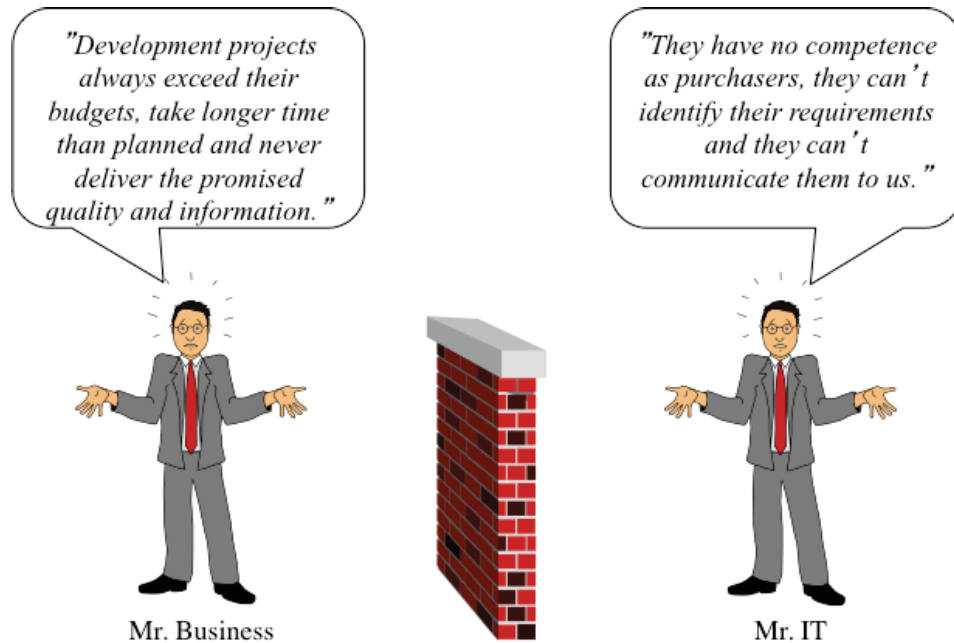


Figure 6. Mr. Business/IT.

defined EA as “the process of translating business vision and strategy into effective enterprise change by creating, communicating and improving the key requirements, principles and models that describe the enterprise’s future state and enable its evolution.”¹⁴ In the United States the government had worked intensively to create a more overarching vision for their work and state that EA “... defines the mission of an agency and describes the technology and information needed to perform that mission, along with descriptions of how the architecture of the organisation should be changed in order to respond to changes in the mission.”¹⁵

The scope for EA is generally widely applicable and possible to adopt in both the public and the private sector. It is possible to create an architecture for an entire business or corporation or a part of a larger enterprise, a conglomerate of several organisations, such as a joint venture or partnership, or a multiply-outsourced business operation. Since the scope could be so wide, an important first step is to define the boundary of the enterprise to be described. “Enterprise” means often more than the information systems employed by an organisation, the term enterprise includes the whole complex, socio-technical system, including people, information, technology and business (e.g., operations). Composing holistic solutions that address the business challenges of the enterprise and support the governance needed to implement them.

4.2 Methods and tools

Enterprise architects use various methods and tools to capture the structure and dynamics of an enterprise. In doing so, they produce taxonomies, diagrams, documents and models. The first attempt to give EA

¹⁴ <http://www.gartner.com/it-glossary/enterprise-architecture-ea/>

¹⁵ <http://us-code.vlex.com/vid/sec-definitions-19256361>

more formal structure was The Zachman Framework for Enterprise Architectures, although self-described as a framework, it is actually more accurately defined as a taxonomy (Figure 7).¹⁶








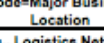




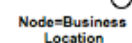

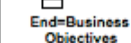
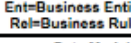
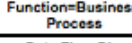
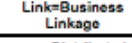
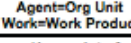
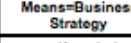


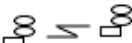
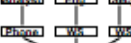
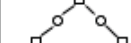
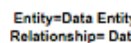

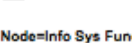

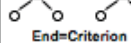
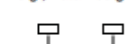

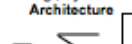
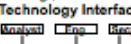
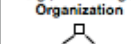
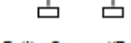
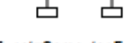


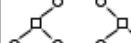
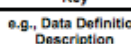
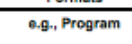
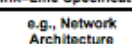
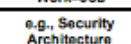
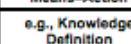

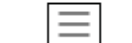


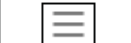
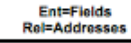
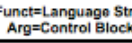
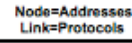
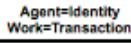
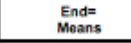
Zachman's Framework for Information Systems Architecture						
	What	How	Where	Who	When	Why
REQUIREMENTS ANALYSIS	Scope	List of Things Important to Business 	List of Processes the Business Performs 	List of Locations Important to Business 	List of Organizations Important to Business 	List of Business Goals/Strategies 
	Investor	Entity=Class of Business Thing 	Function=Class of Business Process 	Node=Major Business Location 	Agent=Class of Agent 	Time=Major Business Event 
	Enterprise Model	e.g., Entity Relationship Diagram 	e.g., Function Flow Diagram 	e.g., Logistics Network 	e.g., Organization Chart 	e.g., Business Plan 
DESIGN	Owner	Ent=Business Entity Rel=Business Rule 	Function=Business Process 	Node=Business Location Link=Business Linkage 	Agent=Org Unit Work=Work Product 	Time=Business Event Cycle=Business Cycle 
	Information System Model	e.g., Data Model 	e.g., Data Flow Diagram 	e.g., Distributed System Architecture 	e.g., Human Interface Structure 	e.g., Knowledge Architecture 
	Designer	Entity=Data Entity Relationship=Data Relationship 	Func=Appl Function Arg=User Views 	Node=Info Sys Funct Link=Line Char 	Agent=Role Work=Job 	Time=Trigger Cycle=Component Cycle 
DEVELOPMENT	Technology Model	e.g., Data Design 	e.g., Structure Chart 	e.g., System Architecture 	e.g., Human/Technology Interface 	e.g., Knowledge Organization 
	Builder	Entity=Segment/Row Relationship=Pointer/Key 	Func=Computer Funct Arg=Screen/Device Formats 	Node=Hardware/System Software Link=Line Specification 	Agent=User Work=Job 	Time=Execute Cycle=Component Cycle 
	Components	e.g., Data Definition Description 	e.g., Program 	e.g., Network Architecture 	e.g., Security Architecture 	e.g., Knowledge Definition 
	Subcontractor	Ent=Fields Rel=Addresses 	Func=Language Strmts Arg=Control Blocks 	Node=Addresses Link=Protocols 	Agent=Identity Work=Transaction 	Time=Interrupt Cycle=Machine Cycle 
	Functioning System	e.g., Data 	e.g., Function 	e.g., Network 	e.g., Organization 	e.g., Strategy 

Figure 7. Zachman grid.

Another, more cooperatively developed framework, which is more accurately labelled as a process is “The Open Group Architectural Framework (TOGAF).” The Gartner Methodology can be best described as an enterprise architectural practice. A blended methodology is that in which bits and pieces from each of these or other methodologies are chosen modified and merged according to the specific needs of your organisation.¹⁷

TOGAF divides enterprise architecture into four categories:

1. Business architecture—Describes the processes the business uses to meet its goals
2. Application architecture—Describes how specific applications are designed and how they interact with each other
3. Data architecture—Describes how the enterprise data stores are organized and accessed

¹⁶ J. A. Zachman, “A framework for information systems architecture,” *IBM Systems Journal* 26 (1987): 276-292.

¹⁷ Sessions, “A Comparison of the Top Four Enterprise-Architecture Methodologies.”

4. Technical architecture—Describes the hardware and software infrastructure that supports applications and their interactions

TOGAF describes itself as a framework. But the most important part of TOGAF is the Architecture Development Method, ADM, a recipe for creating architecture, a process. Viewed as an architectural process, TOGAF in many ways complements Zachman's model which is an architectural taxonomy. Zachman tells you how to categorize your artefacts. TOGAF gives you a process for creating them. TOGAF views the world of enterprise architecture as a continuum of architectures, ranging from highly generic to highly specific. It calls this continuum the Enterprise Continuum. TOGAF's ADM provides a process for driving this movement from the generic to the specific.¹⁸

Examples of enterprises where information management is the business—or an important part of the business:

- Business intelligence systems
- Public libraries, databases, and register systems
- Production of official statistics
- Knowledge bases and open access journals
- Research-support systems
- E-commerce systems
- Archive management

I think it is important to note that most of these enterprises are multi-purpose and serve partly unknown customers and needs. Complex metadata subsystems and data quality issues are typically essential.

There are many different definitions of information architects. In this context, I will mainly focus on the category known as business information architects (Figure 8).

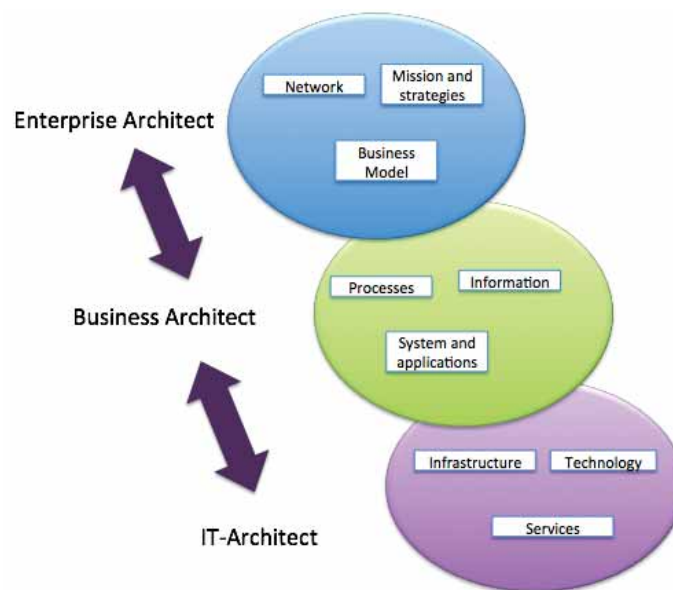


Figure 8. Architect levels.

¹⁸ Ibid.

5. Modeling and analysing the business

The Business Architect works and analyses mainly within these three areas: the processes, the information objects and the IT-system. It's common for the business information architects to start to create a process map (Figure 9).



Figure 9. Modeling and analysing the business.

5.1 Process

The event or events that start the processes are mapped, then the directional flow of processes and the sequence of processes in the business are outlined, finally the outcomes of the processes are documented (Figure 10). The processes are also described in plain text.

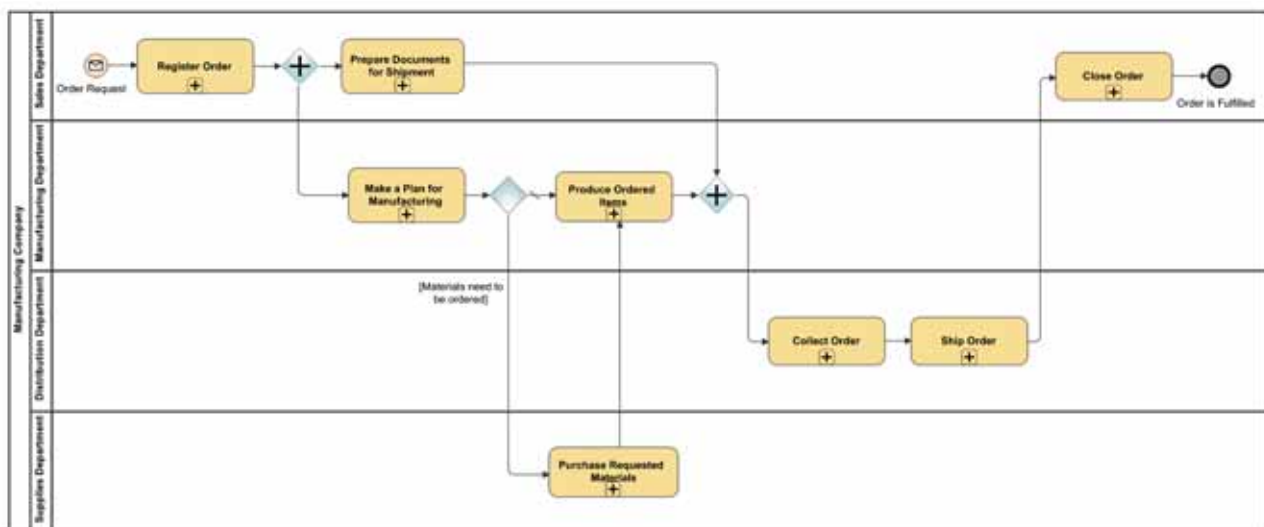


Figure 10. Process Flowchart.

5.2 Information

The next step is to create a comprehensive information model. Master data, that is more stable data, is oriented towards a person, infrastructure or product. Business activities are less stable and have a high modification rate with additional data. An information object must be essential to the business, have one or more attribute and be identified by a stable and unique key (Figure 11).

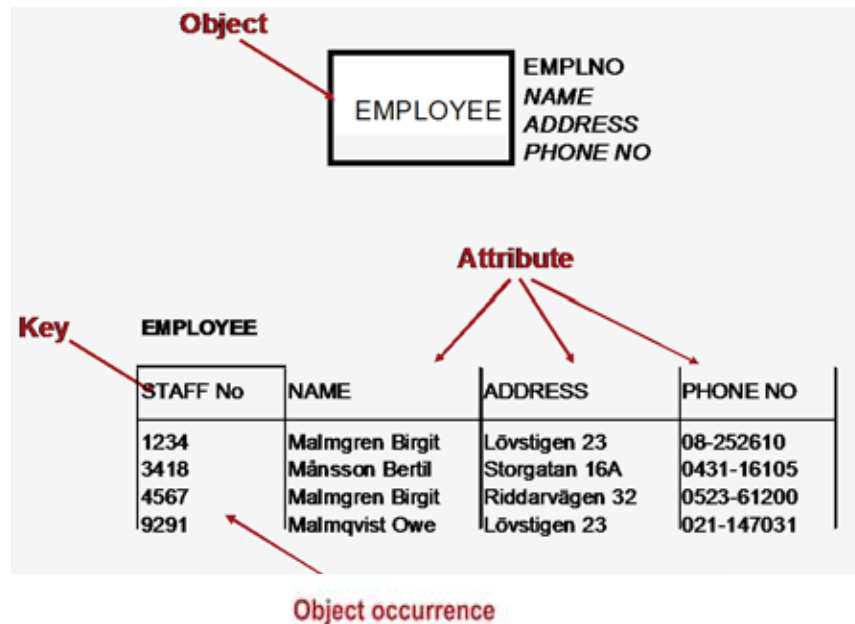


Figure 11. Information objects.

Find the most central information objects—like a classic object as Customer, Employer, and Product, etc.

- Prioritize until a manageable number of key objects remain (about 25-35)
- Make the array of relationships based on these objects
- Have an information modeller make proposals for object grouping. Reconcile this with business, management and IT architects. It is also common to colour-code the different objects/resources, and group them with business events in the middle and the various resources around (Figure 12).

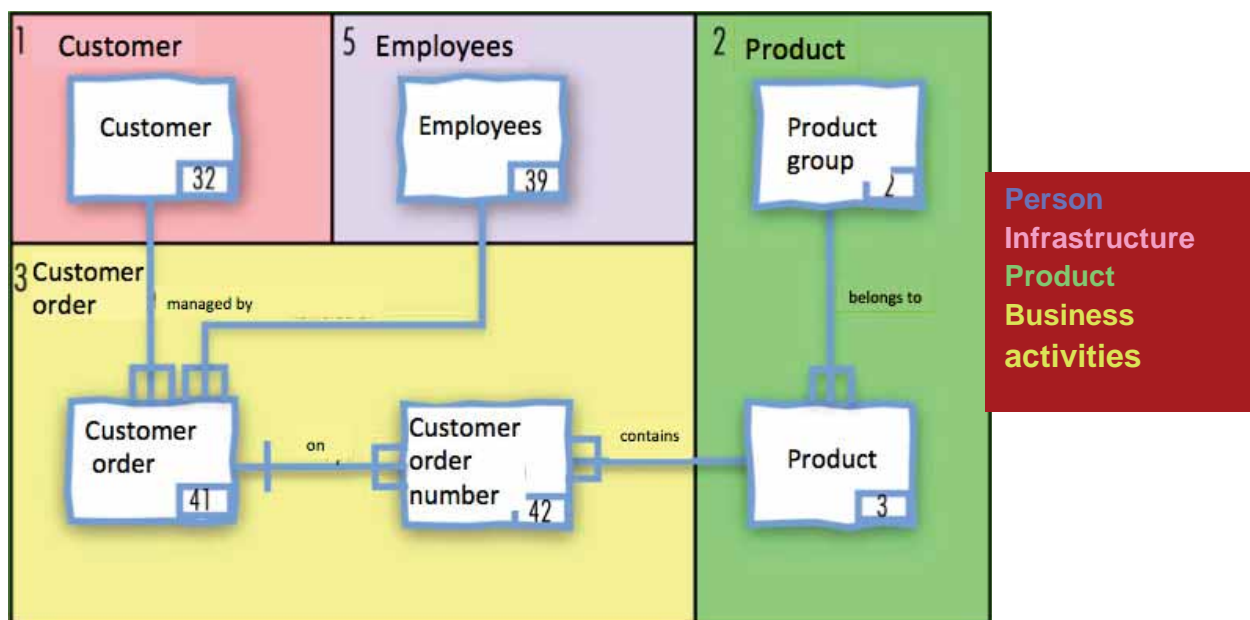


Figure 12. Information objects—summary.

5.3 Analysis

The analysis provided by a business architect is based on the current situation, but its primary purpose is to sketch out a more optimal scenario and improved flow for the business.

To get an overview of how information is handled in the current system a matrix is developed in which the most important IT-systems are represented, indicating which information objects are supported by which system(s) (Figures 13 and 14).

What information is needed in processes?

Process \ Object group	Staff	Customer	Product	Sales order
Create tender	○	●	○	●
Provide service to the customer	○	●	○	●
Manage sales order	○	●	○	●
Purchasing	○	○	○	○

● = Create
○ = Read
⊙ = Update

Figure13. Information analysis.

Process \ Object group	Staff	Customer	Product	Sales order
Create tender	●	○	○	○
Provide service to the customer	○	●	○	○
Manage sales order	○	○	●	○
Purchasing	○	○	○	●

● = Create
○ = Read
⊙ = Update

Figure14. Vision.

Several black dots in the same column mean that the same data is created in multiple processes. It's not good to have different systems that create the same data. Second, it shows that a particular process is dependent on the data produced in another process. This shows the necessity of making data available in

the entire operation to prevent duplication. This provides great potential for improvement. The objective is to:

1. Obtain data once at the source and make it available to all systems
2. Make data/information system structure independent of the organization
3. Ensure that information resources support business development

6. Conclusion - Similarities and Differences

The Business Architect's focus is here and now and then his next step is to create a map showing the way to improvement. The Archivist also focuses on here and now but looks further forward in time—in principle towards eternity. He does not have a stated goal of improving operations as does the business architect. The Archivist's mission is to document the business, as it is—good or bad.

6.1 Common characteristics

The Archivist usually makes a more detailed process map and follows the different parts of the process down to the document/information object level, whereas the business architect usually stays at an overview level. In the two professions' approaches to information modelling one can find buried in the archivist's document plan traces of the business architect's ways of working with information objects. There are, however, still major differences and further work is required to synchronize these two different ways of looking at information objects.

Both groups will benefit greatly by ensuring that their various tools and definitions for process mapping become harmonized, so that both groups use the same scales of notation, definition of common terms, etc. Information modelling can then be supplemented by and harmonized with the archivist's Document Plan

Collaboration is needed—otherwise it will be impossible to achieve the ambition of the regulations in the Swedish National Archive (RA-FS 2009:1). The DIRKS¹⁹ methodology is in many respects similar to developing enterprise architecture, particularly in the analysis of an organisation's business activities (Step C of the DIRKS Methodology). A functional analysis closely matches the business layer of enterprise architecture. Business classification schemes can therefore be very useful resources when shared. Records and archive managers have sound knowledge of their organisation's information assets (taxonomies, retention and disposal authorities, records control systems, preservation regulations).

¹⁹ DIRKS (Developing and Implementing a RecordKeeping System) presented in ISO 15489 is an 8-step methodology familiar to systems developers:

Step A - Preliminary investigation

Step B - Analysis of business activity

Step C - Identification of recordkeeping requirements

Step D - Assessment of existing systems

Step E - Identification of strategies for recordkeeping

Step F - Design of a recordkeeping system

Step G - Implementation of a recordkeeping system

Step H - Post implementation review

Read more here: <https://www.records.nsw.gov.au/recordkeeping/dirks-manual/introducing-the-dirks-methodology/dirks-methodology-and-manual>.

Business architecture is also used as a strategic tool, aiming to influence how the organization works towards achieving its goals/mission. Linking archival and records management goals to that vision of an organisation's future promotes the strategic development of the organisation's archive- and records management capacity.

In today's environment in which an increasing volume of information is used by more and more over longer periods of time, the need for orderliness and collaboration is increased. Thus it is very important that the long-term accessibility of information is considered in the initial mapping of business activities. If we want to create better quality in our information flows, it is necessary that the two groups start to cooperate. So it's time to create benefits for the future—begin immediately to integrate the long-term supply of information in the information architect's first sketch map. Finally, I would like to emphasize the importance of including the archivist clearly and highlight their role in the future of information management and the need of a common map. If you do not know the past, you will not understand the present, and then not be able to shape the future.

Enterprise Content Management and Digital Curation Applications

Maturity Model Connections

Shadrack Katuu

International Monetary Fund

Abstract

Organizations have a variety of business systems to help them manage their digital content. Depending on the institution, the digital content connected to these systems could either be managed in network drives, or through the use of specialized business applications including ECM applications or even left unmanaged. This paper narrowly focuses on the issues related to the transfer of digital content from Enterprise Content Management (ECM) applications to Digital Curation (DC) applications. It does so by defining ECM applications and their relations with other similar applications such as Electronic Document Management Systems and Electronic Records Management Systems. It also defines Digital Curation applications and highlights why they are different from ECM applications. The paper argues that the process of transfer digital content from ECM to DC applications requires support of maturity models. At the core of maturity models is the quest for continuous improvement by providing a framework of assessing processes and a roadmap for advancement. The article provides an outline of one model for ECM applications and another related to digital preservation, illustrating briefly their utility. It concludes by stating that there are few published articles on the transfer of digital content from ECM to DC applications and theirs is yet more insight to be gathered from the few examples so far. It adds that maturity models are not in themselves perfect and continuously need improvement, which is at the core of the mission of maturity models.

Author

Shadrack Katuu has had a diverse international career, spanning various information management fields in several countries including Botswana, Canada, Kenya, Namibia, South Africa, Swaziland and the US. He has an undergraduate degree from Kenya and studied at the School of Library, Archival and Information Studies at the University of British Columbia in Canada. He is currently undertaking doctoral studies with the University of South Africa. He was a staff member at the Nelson Mandela Foundation (NMF) between 2005 and 2009, and for the latter half of the duration was manager responsible for all information systems including the institution's ECM application. He is currently an archives/records officer at the International Monetary Fund (IMF) in Washington DC.

1. Introduction

Organizations have a variety of business systems to help them manage their digital content. These business systems are connected to a number of functions and activities including human resources (such as recruitment and payroll), communication (through using email), finance, marketing and other aspects of administration. Depending on the institution, the digital content connected to these systems could either be managed in network drives, or through the use of specialized business applications including ECM applications or even left unmanaged.

The way these systems are connected within an institution can be quite complex. The diagram below provides an illustration of how some of the business systems were organized in 2003 at the World Bank Group (Van Garderen 2002).

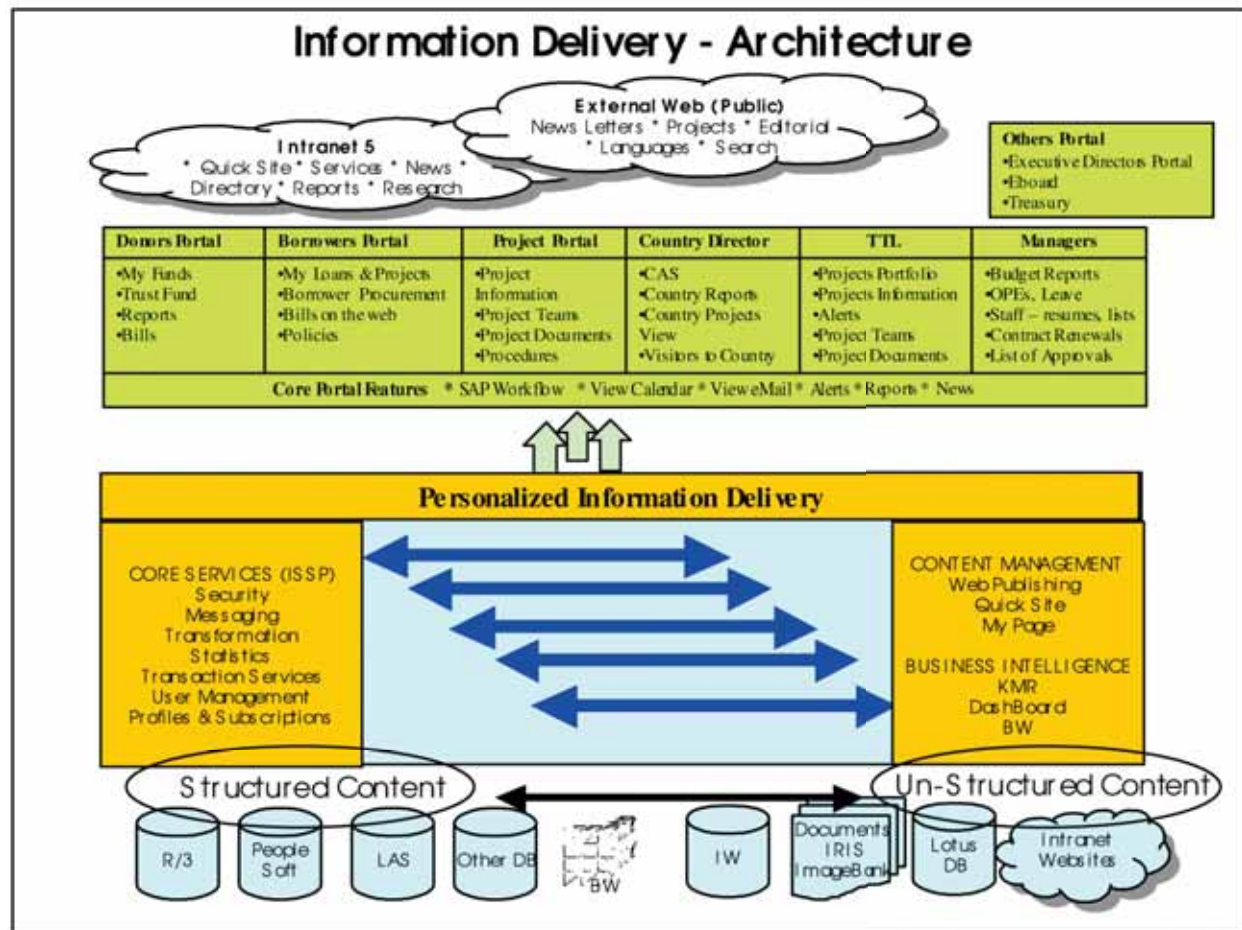


Figure 1. Showing the information delivery architecture at the World Bank.

The diagram above shows that in order for the World Bank to conduct its business, it depended on numerous applications. For purposes of accountability, it would be critical to keep the information generated and maintained in these applications over the long-term. Most business systems are, however, not designed to keep transaction information in the long-term, so for this reason, a different set of applications, Digital Curation systems, could be considered. This is not only true for the World Bank but for other institutions such as municipal governments like the City of Vancouver in Canada (Dingwall 2011) or an international criminal court like the International Criminal Tribunal for Rwanda (Peterson 2008).

Whenever organizations consider managing digital content in the long-term, moving the content into Digital Curation systems becomes a challenge. This paper narrowly focuses on the issues related to the transfer of digital content from Enterprise Content Management (ECM) applications to Digital Curation systems.

Defining ECM

Enterprise Content Management (ECM) is a concept that has been used by information professionals for more than a decade. As early as 2001, Karen Shegda from Gartner, a leading research and advisory firm, discussed integrated document management software functionality and noted that software vendors were

morphing their products into content management systems (Shegda 2001). In the same year Bob Ward of Teamware Group, a technology consulting company, published an article arguing that content management was a growing sector in the information technology industry and titled the article enterprise content management (Ward 2001).

For a long time the term ECM has been used interchangeably with electronic document and records management systems and other concepts. For the purpose of this article, ECM is viewed currently as the most sophisticated point in an evolutionary process. The other predecessor points in this evolutionary point are Electronic Document Management Systems (EDMS), Electronic Records Management Systems (ERMS), Integrated Document and Records Management Systems (IDRMS) and Electronic Document and Records Management Systems (EDRMS). This evolutionary perspective accommodates predecessor concepts (Sprehe 2005) and would help clear any confusion regarding the different concepts (Nguyen L T, Swatman P M C, and Fraunholz 2007). The evolutionary process is illustrated in Figure 2.

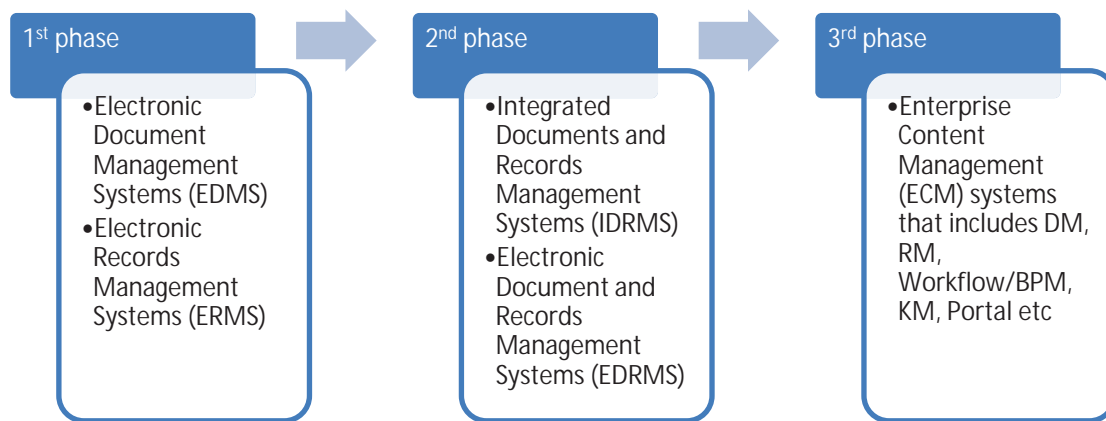


Figure 2. The evolution of various concepts culminating into ECM.

This evolutionary perspective to the concept of ECM is supported by published reports from leading research organizations in document and records management. These reports have, over the last few years, evolved from using terms such as IDMS and EDMS to ECM. Gartner published a report in 2003 that used the concept IDMS (Gartner 2003), but from 2004 used ECM as a concept (Shegda et al. 2004). Another leading research and advisory firm, Forrester, had already used the term ECM in a report published in 2003 (Moore and Markham 2003) and continued to use the term in subsequent annual reports.

In order to sufficiently differentiate ECM with predecessor concepts it is important to define it and provide constituent parts. AIIM (2010) defines ECM as constituting “strategies, methods and tools used to capture, manage, store, preserve and deliver content and documents related to organizational processes.” ECM applications and strategies allows the organization to manage its information more effectively (AIIM 2010).

When these strategies, methods and tools are targeted at organizational processes, they manifest themselves in several modules. The precise number and composition of the modules remains a subject of debate. For the purpose of this article, the 10 modules considered fundamental include: Document Management (DM), Records Management (RM), Workflow or Business Process Management (BPM), Collaboration, Portal, Knowledge Management (KM), Imaging, Digital Asset Management (DAM),

Digital Rights Management (DRM), and Web Content Management (CMS Watch 2010, 21-86; Kampffmeyer 2004, 2006). Figure 3 shows a graphical representation of these modules of ECM.

For the most part, scholarly discussions on the systematic management of digital content have dwelt on EDRM applications (Wilkins, Swatman, and Holt 2009; Wilhelm 2009; Biagio and Ibricu 2008; Scott-Jones 2002). Based on the foundation laid in this article, EDRM applications have merely two of the modules that would be available in the ECM suite. In a survey of 10 South African institutions that considered themselves as having implemented ECM applications, nine of the institutions had both records and document management modules (Katu 2012b). If these were the only modules implemented, then it would mean EDRM applications are in place. However, most of these institutions had more than just DM and RM modules in place which demonstrates that institutions tend to go beyond just two ECM modules (Katu 2012b, 50). These realities are important to bear in mind when considering the preservation of digital content in the long-term. This is because the more modules an ECM application has the more complex the transfer process tends to be.

1.2 Defining Digital Curation applications

There have been debates about the differences between digital preservation and digital curation (Lazorchak 2011). Elizabeth Yakel argues that digital preservation is a subset of digital curation which

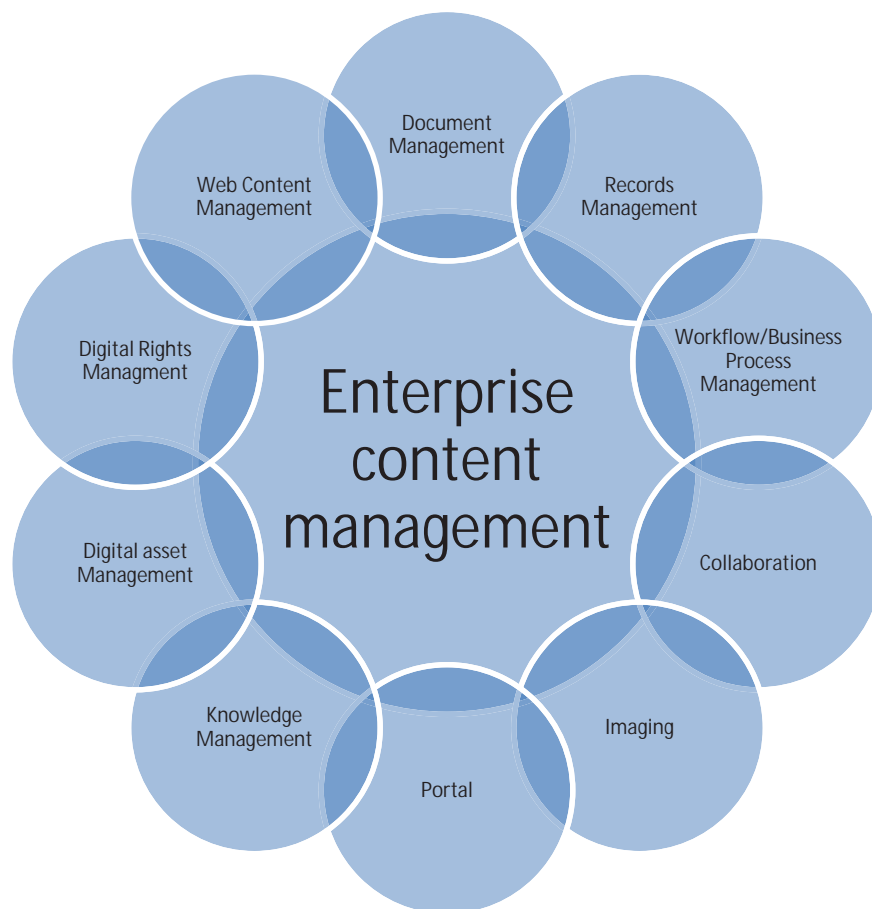


Figure 3. The modules of a typical ECM application.

she defines as “the active involvement of information professionals in the management, including the preservation, of digital data for future use” (Yakel 2007). According to Foscarini, Kim et al. (2010, 1) digital preservation “aims to ensure that digital objects of value to society...can be meaningfully reproduced over time, despite evolving representations, mechanisms, rapidly advancing technologies, and continually emerging user expectations.” While digital curation is holistic and spans the whole of the lifecycle of the digital content, digital preservation is a specific point in that lifecycle (Caplan 2011).

Lee and Tibbo (2011, 126) trace the emergence of the concept of curation from diverse perspectives and how it has served as “an umbrella concept spanning activities across a diversity of professions, institutions..and sectors.” At the core of digital curation activities are digital repositories. A digital repository is seen as “a combination of services, resources that are required to carry out those services as well as supported by the service, and policies that determine how the services should be implemented” (Lee et al. 2009, 113). They add that no two repositories regardless of configuration or use of same kinds of infrastructure will be the same since they will differ based on services, resources or policies (Lee et al. 2009, 113).

Digital repositories range from very large storage management schemes, peer-to-peer integrity checking among mirrored repositories (Galloway 2009, 1519) as well as micro-services based repositories (Archivematica 2012). These technologies have adhered to the OAIS reference model which has become a basic reference point for discussions on digital repositories (Galloway 2009, 1521). These technologies could also be grouped according to whether they are open source or proprietary systems.

2. Transfer of digital content from ECM applications to Digital Curation applications

There has been an argument that if digital content is already in ECM applications then those applications should be used as Digital Curation systems. According to Seles (2012), EDRMS applications could be used as holding tanks if institutions cannot support digital repositories. Indeed, in some instances, ECM applications have been modified to become Digital Curation systems. For example, the State of Victoria in Australia commissioned a digital repository in 2005 that used an ECM application (Waugh 2007). Another example is the Municipality of Lisbon in Portugal that was reportedly integrating an ECM application “with a set of business workflows for a wide range of organizational entities, including the Municipal Archives” (Becker et al. 2011, 8)

However, according to Waugh (2007), the fundamental operational model of a commercial ECM application is different from that of a digital repository and, therefore, in Australia’s experience, this reality caused several challenges. This sentiment is echoed also by the City of Vancouver Archives in Canada which sought to use a dedicated Digital Curation system to import digital records from their ERDMS application (Dingwall 2011). The issue of incompatibility stems from the fact that ECM applications and Digital Curation applications have drastically different functions and, in a standards based approach, would have to meet different functional requirement guidelines.

On the one hand, there are a diverse number of functional requirements, standards, and best-practice guidelines that relate to ECM applications. These range from those within certain national jurisdictions such as Australia’s ERMS functional requirements (National Archives of Australia 2007) and the United States’ DOD standard 5012.5 whose latest version is version 3 (Department of Defense [United States] 2012) to regional ones such as European Union’s MOREQ standards whose latest version is MOREQ2010 (DLM Forum Foundation 2011). The International Council on Archives (ICA) developed a high-level and generic set of functional requirements for ERMS applications (International

Council on Archives 2008, 8) that seeks to serve a more global audience. According to the ICA functional requirements, there are eight functional requirements: capture, identification, classification, managing access and security, managing hybrid records, retention and disposal, administration and finally search, retrieval and rendering (International Council on Archives 2008).

On the other hand, Digital Curation applications that adhere to the OAIS reference model have seven main functions: access, administration, archival storage, common services, data management, ingest, and preservation planning (Lee 2009, 4025). The City of Vancouver has made this argument from a graphical perspective as represented in the illustration below (Van Garderen 2012).

As Figure 4 shows, while the ERDMS may be able to provide access to its own records, these would only be available to the City of Vancouver staff and not members of the public. In addition, it would be incapable of processing digital content from diverse sources such as other business systems. This is because in most cases, as Mumma, Dingwall et al. (2011, 118) argue, digital content from diverse sources does not come in “neat packages ready for Ingest.” In addition, there are number of archival processes that digital content needs to undergo. In traditional settings, archival processing constitutes accessioning, appraisal, arrangement and description and the provision of access to archival material (Pearce-Moses 2005; Spiro 2009). For the Archives of the International Monetary Fund (IMF), these basic activities take a slight sophistication with two appraisal processes and an additional process for review of confidential material through declassification, as shown in Figure 5 (International Monetary Fund 2012).

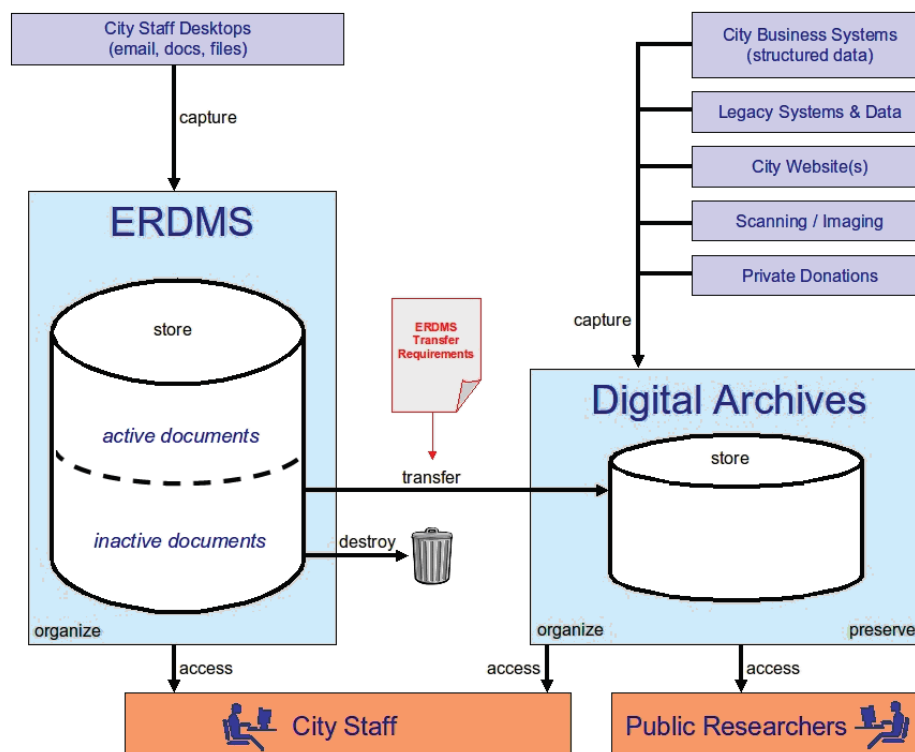


Figure 4. Showing the transfer process at the City of Vancouver Archives.

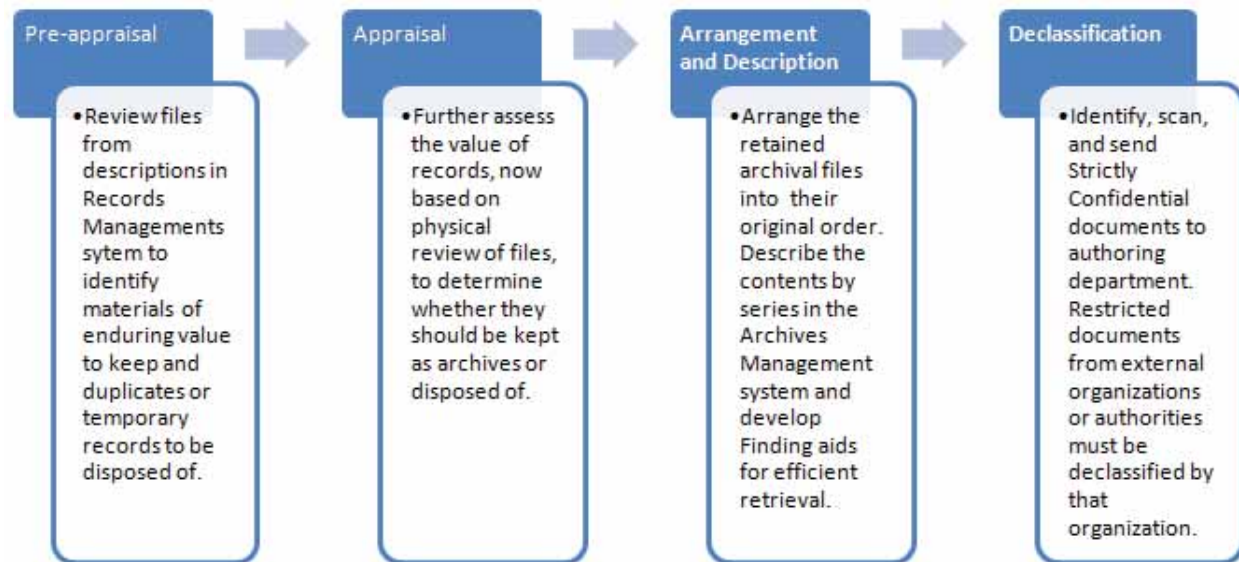


Figure 5. Showing the archival processes undertaken by the Archives of the International Monetary Fund.

When processing digital records of the Vancouver Organizing Committee for the 2010 Olympic and Paralympic Winter Games, the City of Vancouver Archives identified three distinct stages of appraisal: Selection for Acquisition, Selection for Submission and Selection for Preservation (Mumma, Dingwall, and Bigelow 2011, 108). Presumably the first and third stages would mirror the two stages at the IMF Archives while the second stage emerges due to the realities of digital aspects of the records.

Based on these two examples, it would be very difficult to expect ECM applications to provide OAIS functionality such as being able to ingest and manage digital content from diverse record sources and/or providing public access to digital content (Dingwall 2011).

3. Enhancing reliability of digital content using maturity models

As digital content is being transferred from ECM applications to DC applications, there is concern about the trustworthiness of the records. Potentially, this could damage the trustworthiness of the digital content over time. During the first phase of the InterPARES research project, which examined attributes of digital records,¹ researchers found that often, not all these attributes were captured during the transfer process (MacNeil 2002, 53). One such attribute of digital records is the archival bond which is defined as “the relationship that links each record to the previous and subsequent one and to all those which participate in the same activity” (Duranti and MacNeil 1996, 49). In an ECM application, the archival bond may manifest itself as “a classification code or other record identifier that appears on the face of the record or

¹ According to the InterPARES Project, digital records possess a number of characteristics including a *fixed documentary form*, a *stable content*, an *archival bond with other records*, and an identifiable juridical-administrative, provenancial, administrative, procedural, documentary and technological context. They participate in or support an *action*, and at least three *persons* (author, writer, and addressee) are involved in their creation (MacNeil 2002, 29).

in its profile” (MacNeil 2002, 29). The challenge in the transfer process is to ensure that the archival bond is still identifiable. In a research study of the City of Vancouver’s Geographic Information System, Dingwall, Marciano et al (2007, 186-188) illustrated the preservation information that would be needed in order to create the archival bond which would then be revealed through the archival description process. This is just one example of one characteristic but illustrates the need to “ensure that knowledge of key indicators of identity is not lost when the records are removed from the specific electronic system and record-keeping environment in which they have been created and actively used” (MacNeil 2002, 33).

There are different aspects to the trustworthiness of records. One such aspect is reliability and is defined as the trustworthiness of a record as a statement of fact and is established by examining the completeness of the record’s form² and the amount of control exercised on the process of its creation (InterPARES 2 Project 2007).

For purposes of this discussion, the process of record creation is most significant. MacNeil (2000, 101) defines this process as constituting a “body of rules governing the making, receiving, and setting aside of records. Some of these rules refer to record-makers by establishing who is competent to create, modify, and annotate records. Others refer to how records must be handled in the course of their compilation, and others still refer to how records must be routed and filed.” Regardless of the software applications, the rules on establishing who is competent to create, modify, and annotate records should be established using workflow procedures as well as access privilege management (MacNeil 2000, 101). In addition, there should be ways of verifying if such rules have been followed using audit trails which entails “recording all the interactions with records within a system so that any access to the record can be documented as it occurs” (MacNeil 2000, 102). According to Duranti (1995, 6), “[t]he more rigorous and detailed the rules, the more established the routine, the more reliable the records result from the application will be.”

One concept that could be used to establish routines as well as improve on rigor and details of rules of both ECM and DC applications is maturity models. A maturity model is a management tool designed to help organizations implement effective processes in a given management discipline. It is a “structured collection of elements that describe characteristics of effective processes.” It provides a place to start, the benefit of prior experience, a common language, a framework for prioritizing actions and a way to define improvement” (Murray and Ward 2007, 5).

Even though the concept has existed for almost three decades (Cameron 2011, 21), very little has been published in connection with ECM and DC applications. The most prominent early application of maturity models is in computer software engineering in the 1980s and 1990s (Liu 2002) which later spread to other disciplines including, financial management (McRoberts and Sloan 1998), human resources management (Curtis, Hefley, and Miller 1995), the health sector (Gillies 2000), and project management (Kerzner 2001).

Lee and Tibbo (2011, 127) argue that the challenges associated with digital curation are not solely technical. According to Dryden (2011, 128), trust is critical when dealing with a digital repository and asks “[h]ow would one know that a digital repository could preserve digital information so it would be accessible over time for as long as it is needed?” To this end, a number of efforts have taken place around the world to develop audit mechanisms for repositories. These include the German NESTOR³ project that

² According to MacNeil (2000, 100) a complete record is one that that posses “all the elements of intellectual form necessary for it to be capable of generating consequences”. These elements include: “date of record; time and place of creation, transmission, and receipt; identification of names of author, addressee, originator and writer (if either or both are different from the author), name (or crest) of creator, title or subject line, classification code, and any other element required by the creator’s procedures and/or juridical system.”

³ This is the acronym for “Network of Expertise in Long-term Storage of Digital Resources.”

developed a *Catalog of Criteria for Trusted Digital Repositories*, as well as the UK and Netherland DRAMBORA project that developed a risk based audit mechanism (Dryden 2011, 128-129). The International Standards organization published a standard in 2012 titled *Audit and Certification of Trustworthy Digital Repositories* that draws from a lot of the previous work done around the world in order to “enable the assessment and certification of a repository as being a trustworthy digital repository (TDR)” (Kroll et al. 2012). According to Cho (2012), while ISO 16363 provides for compliance requirements it does not provide for how to control performance of digital repositories and how to improve organizational capabilities over time. Therefore, Cho (2012) proposes the use of maturity models to address this shortfall.

Regardless of their disciplinary application, maturity models are developed on the basis that organizations do not move from zero capability to optimum capability instantaneously, but rather progress along a journey of maturity (Murray and Ward 2007, 5). This journey of maturity is documented in a number of levels of maturity. The number of levels may vary from three to six but five and six are the most common levels. Sections 3.1 and 3.2 below will outline a number of maturity models that have been proposed for ECM and DC applications respectively.

3.1 Maturity model for ECM

There are at least three ECM maturity models that have been developed to date. One was developed by an institution in South Africa (Katuu 2012a) while another was developed in the UK (Cameron 2011). Neither of these models have received much global attention. The third model, known as the ECM Maturity Model (ECM3), is probably the most well known and was developed in the US by four consulting firms as an open standard under creative commons license (MIKE2.0 2010). This section will highlight aspects of ECM3.

The first edition of ECM3 was published in March 2009 and a second edition in March 2010 (Pelz-Sharpe et al. 2010). ECM3 provides a structured framework from which to organize efforts by organizations to achieve business benefits from ECM, as well as to hold the attention of program stakeholders. (Pelz-Sharpe et al. 2010, 7). It is able to do this because it can be applied to audit, assess, and explain the current state within an organization, as well as form a roadmap for maturing organization capabilities (MIKE2.0 2010). The framework has 13 dimensions of maturity across three categories: human, information, and systems (Pelz-Sharpe et al. 2010, 8). The dimensions within the “Human” category relate to individual expertise in both business processes and information technology. In addition, they relate to the extent of strategic alignment between business drivers and the ECM application to ensure institutional success. The dimensions within the “Information” category relate to attributes affecting the digital content itself, while those within the “Systems” category relate to attributes of the ECM applications technical features. Figure 6 provides a graphical representation of the maturity dimensions across three categories (MIKE2.0 2010).

The three categories and 13 dimensions are assessed in the five levels of maturity below:

- Level 1: Unmanaged
- Level 2: Incipient
- Level 3: Formative
- Level 4: Operational
- Level 5: Pro-Active (Pelz-Sharpe et al. 2010, 6).

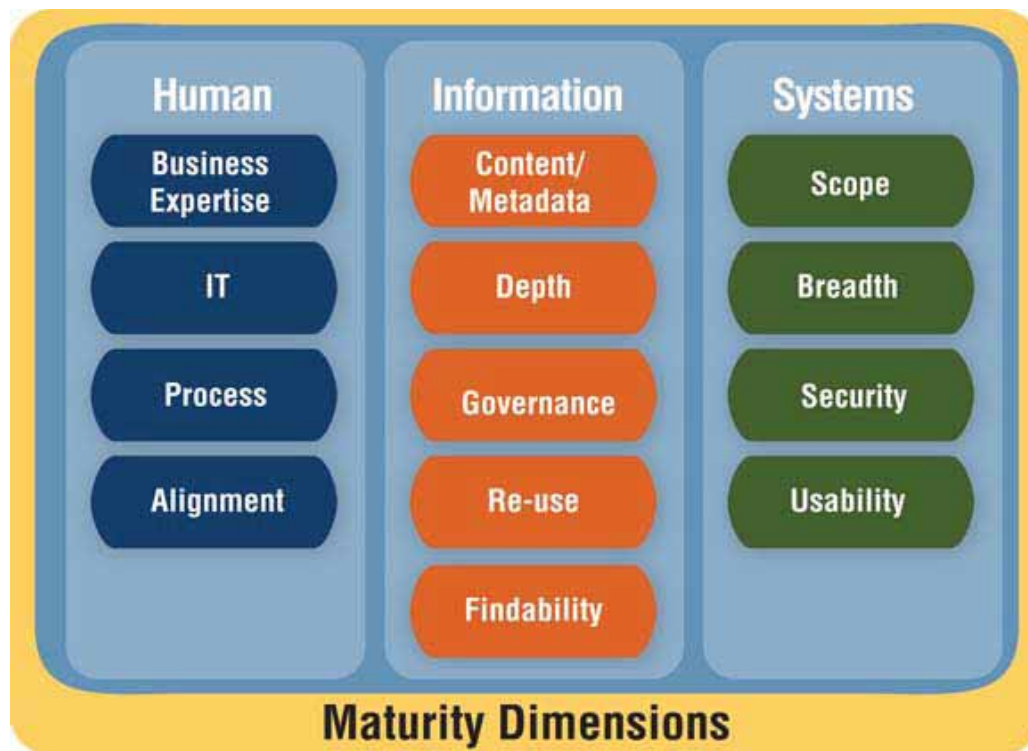


Figure 6. Showing the maturity dimensions in the ECM maturity model.

When the levels of maturity are combined with the 13 dimensions, a comprehensive chart is developed, as illustrated in Figure 7 (MIKE2.0 2010). This diagram provides a sample assessment of a hypothetical Enterprise ABC showing different maturity levels achieved for each individual dimension (Pelz-Sharpe et al. 2010). Even though the ECM Maturity Model is reportedly used by a large number of institutions (ECM Maturity Model 2012) there has been little published about the experiences within these institutions (Katu 2012a).

3.2 Maturity model for digital preservation and digital repositories

There are at least three maturity models that relate to either digital preservation or digital repositories. These are: *Digital Preservation Capability Maturity Model (DPCMM)* (Dollar and Ashley 2012), *Shaman/Scape Capability Model* (Becker et al. 2011) and the *Trusted Digital Repository Maturity Model (TDRMM)* (Cho 2012). One notes that all these models are at different stages of development and have different areas of emphasis due to their different perspectives. This is a reflection of how new the concept is in this discipline. The first has been developed to the point of being used in institutions while the second and third are still being developed. This section will only highlight the first model.

According to Dollar and Ashley (2011, 8), the objective of DPCMM is to “provide a process and performance framework...against best practice standards and foundational principles of records management, information governance, and archival science.” DPCMM draws from the functional specifications and preservation services identified in ISO 1472, as well as checklist criteria found in TRAC guidelines (Trustworthy Repositories Audit and Certification: Criteria and Checklist). It is centered a trusted digital repository which is the results of digital preservation infrastructure as well as

← “Enterprise ABC” ECM Maturity Levels

Level: 1) Unmanaged		2) Incipient	3) Formative	4) Operational	5) Proactive
Dimension:					
HUMAN	IT Expertise	No experience managing formal repository and workflow systems	Struggling 1.0 implementations of some systems	More advanced version 2.0+ implementations of systems, with focus on business-critical content	Managing repository and workflow systems is a core IT skill
	Business Expertise	Ignorance about value and role of ECM	Growing sense of awareness about lack of management services	Communication plans include updates to key stakeholders about ECM business value	Executive sponsorship of ECM as a practice; process and content analysis are core skills
	Process	Few or no standardized procedures around content	Basic process analysis leads to some ad-hoc workflows	Initial modeling of inter-departmental processes to prep for automation	Automated processes span systems and departments
	Alignment	Key business drivers are not well understood by IT strategists, resulting in ECM gaps in IT portfolio	Gaps still exist between technology and core business processes; IT-metrics not evaluated by business outcomes	IT and Business both understand their information management roles and their respective strategies are no longer developed in a vacuum	Execution of IT & Business strategies become more cohesive, but still follow push-pull model
INFORMATION	Content/metadata	No formal inventory; no formal classification	Departmental inventories and initial content tagging	Enterprise inventory underway; controlled vocabularies (CVs) initiated	All new repositories and content types registered; global taxonomies created
	Depth	No lifecycle management	Most content archived haphazardly; some departmental RM efforts	Development of formal electronic retention, RM, and disposition schemes	Implementation of electronic and paper-based RM across the enterprise
	Governance	No policies and procedures	Scattered policies; few or no formal procedures	Development of information governance structure and codification of procedures	Policies and procedures widely disseminated; Enterprise ownership in place
	Re-use	Content routinely duplicated	Content still routinely duplicated	Initial content analysis and structuring	Documents repurposed across systems and channels
	Findability	Employees spend excessive time searching using various internal search engines	Search indexes tuned and basic metadata applied	Rationalization of search technology; analysis of search logs and further tuning, leveraging CV terms	Development of specific enterprise and/or federated search applications
SYSTEMS	Scope	No understanding of core content types	Some basic DM implementations with ad hoc workflow	Identification of core content types, locales; pilot projects for DAM, BPM, etc.	Business-critical information systems prioritized
	Breadth	No systems	Scattered departmental efforts	Initial attempts to combine or integrate systems across departments	Successful departmental initiatives have been scaled enterprise-wide
	Security	No security regime in place	Dependent on individual systems	Formal projects initiated to address gaps & redundancies due to multiple solutions	Standardized policies and procedures exist and are system enabled
	Usability	Lack of systems make end user usability considerations moot	Employee adoption rates measured, but dissatisfaction unanalyzed	Some initiatives use Scenario Analysis and User Persona techniques to guide design	User-centered design underpins all system designs, with formal collection of user feedback

Measurement / Monitoring and Feedback Processes

Figure 7. Showing an outline of the levels and dimensions of maturity in the ECM Maturity Model.

digital preservation processes (Dollar and Ashley 2012). Figure 8 is a graphical representation of the model (Dollar and Ashley 2012).

The model has 15 process elements that are assessed using five maturity levels. According to Dollar and Ashley (2012). Amongst the strengths of DPCMM is that it shows clearly defined components which then enables priority setting based on risk, requirements, and resources for digital preservation processes. Figure 9 illustrates the kind of roadmap that could be developed based on an assessment of an organization using the model and demonstrates how the model could be used to monitor incremental improvements over time (Dollar and Ashley 2012).

The model has been used by a number of institutions, primarily in the US and Canada. These include: Council of State Archivists (Corridan 2012), City of Toronto (Melikhova 2010), Delaware State Library and Archives, and Georgia State Archives (Dollar and Ashley 2012).

4. Conclusion

This article began by stating that organizations use a variety of business systems to assist in carrying out their mandate and that ECM applications are just one type. However, the focus on ECM applications was not just about managing the scope for discussion but also because surveys have shown that organizations

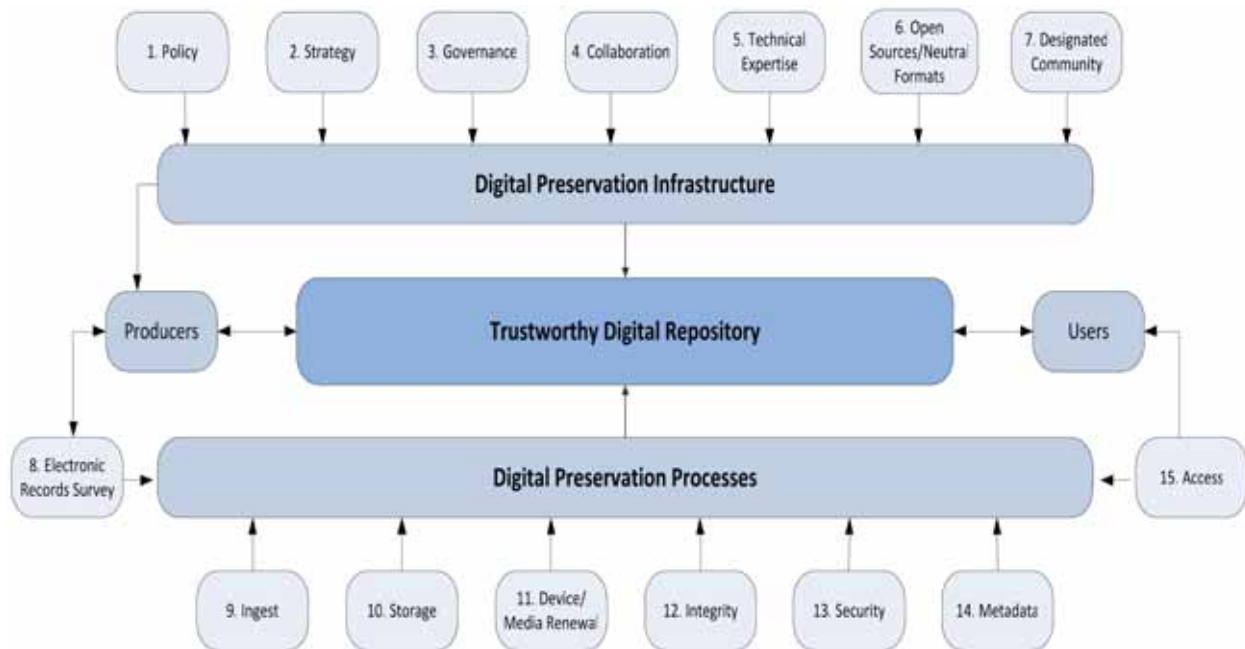


Figure 8. Digital Preservation Capability Maturity Model.

	Current	Year 1	Year 2	Year 3	Year 4	Year 5	Difficulty
Reference Model Components							
PROCESSES							
Electronic Records Survey	1	⇒	2	3	⇒	⇒	MEDIUM
Ingest	0	⇒	⇒	1	⇒	2	HIGH
Archival Storage	0	⇒	⇒	1	2	⇒	LOW
Media & Device Renewal	1	⇒	⇒	2	⇒	3	MEDIUM
Integrity	0	⇒	⇒	⇒	1	2	MEDIUM
Security	1	⇒	⇒	2	⇒	3	MEDIUM
Preservation Metadata	0	⇒	⇒	1	2	3	MEDIUM
Access	1	⇒	⇒	⇒	2	⇒	MEDIUM

Figure 9. Showing Digital Preservation Capability Improvement Road Map.

are increasingly adopting these applications to manage their digital content. Over the last three years, AIIM has conducted a number of surveys on the state of the ECM industry in the world. In 2009, 43% out of a total of 478 institutions from five continents were either implementing or had completed implementation of ECM applications (Miles 2009, 7). In 2010 the rate was at 38% out of a total of 680 institutions worldwide (Miles 2010, 7) and in 2011 the rate was at 35% out of a total of 586 institutions worldwide (Miles 2011, 8). When one adds the number of institutions that were planning to implement these applications, then the figures go well beyond 50%.

This article has made the argument that ECM and DC applications fulfill different functions and, therefore, should be considered as separate systems. Regardless, there's a requirement that they don't compromise the trustworthiness of the digital content that they manage. In order to maintain and improve reliability, rigor is required in the process. Maturity models have been proposed as a means of improvement and the diagram in Figure 10 provides an illustration.

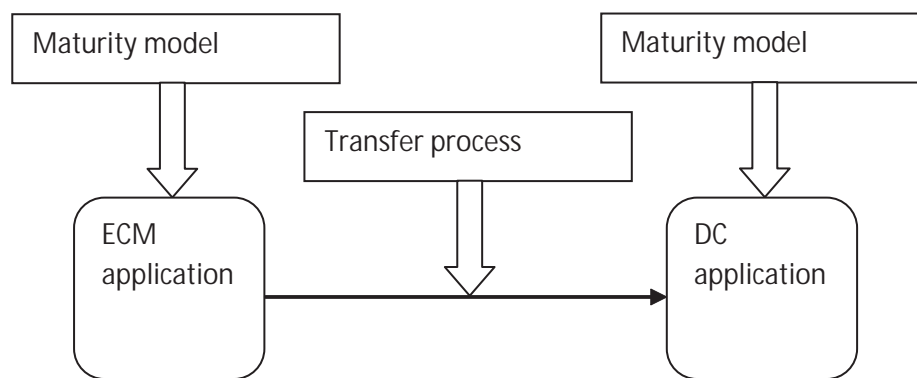


Figure 10. Showing maturity model's influence on ECM and DC applications.

To the best of this author's knowledge, not many institutions have extensive experience in the transfer process from ECM applications to DC applications. The City of Vancouver Archives has considerable experience processing records of the 2010 Winter Olympic games (Mumma, Dingwall, and Bigelow 2011), while the IMF Archives has some experience in the processing of email records (Jordan 2011). The City of Vancouver is expected to start transferring records from its ERDMS in 2012 (Mumma, Dingwall, and Bigelow 2011, 95) and, therefore, the lessons learned from this process may yet take a little while longer before they are shared in professional forums.⁴

Therefore, the suggestion being made in this article for the incorporation of maturity models in improving both ECM and Digital Curation systems may be, from a practical perspective, a bit far-fetched since institutions have yet to have extensive experience in the transfer process. In addition, while maturity models may add value they are, for the most part, still in the early stages of their development and only through extensive use would they be able to provide opportunities for further refinement. Cho (2012) states that most maturity models are structured either as grid-based or stage-based which essentially means that their perspectives are two dimensional. He argues that there's need to combine both perspectives in order to enhance the models (Cho 2012).

⁴ Aspects of the lessons to be learned are being shared on the Artefactual wiki http://artefactual.com/wiki/index.php?title=VanDocs_Interface but there is very little information.

References

- AIIM. *What is Enterprise Content Management?* 2010. Accessed 5 March 2011.
<http://www.aiim.org/What-is-ECM-Enterprise-Content-Management.aspx>.
- Archivematica. *Micro-services*. Archivematica, 2012. Accessed 2 August 2012.
<https://www.archivematica.org/wiki/Micro-services>.
- Becker, Christoph, Gonçalo Antunes, José Barateiro, and Ricardo Vieira. *Schaman/Scape Capability Model*. National Library Board Singapore and Nanyang Technological University, 2011.
http://www.ifs.tuwien.ac.at/~becker/pubs/becker_ipres2011.pdf.
- Biagio, Maria Luisa Di, and Bernice Ibiricu. "A balancing act: learning lessons and adapting approaches whilst rolling out an EDRMS." *Records Management Journal* 18, no. 3 (2008): 170-179.
- Cameron, Stephen A. *Enterprise content management - a business and technical guide*. Swindon, UK: The Chartered Institute for IT, 2011.
- Caplan, Priscilla. *Re:[digital-curation] Semantics: Digital Preservation vs. Digital Curation*, 2011. Accessed 20 September 2012. <https://groups.google.com/forum/?fromgroups=#!topic/digital-curation/ehppkZT9XGs>.
- Cho, Namdo. *Trusted Digital Repositories Maturity Model (TDR-MM)*, 2012. Accessed 17 September 2012. <http://aeri2012.wordpress.com/conference-schedule/paper-presentations/trusted-digital-archives/>.
- CMS Watch. *The ECM Report 2010*, 2010. Accessed 25 March 2011.
<http://replay.waybackmachine.org/20090524222117/http://www.cmswatch.com/ECM/Report/>
- Corridan, Jim. *State Electronic Records Initiative: An Update*. 2012. Accessed 8 August 2012.
<http://www.digitalpreservation.gov/meetings/documents/ndiipp12/Sessions/Preserving%20Electronic%20Records%20in%20the%20States/SERI%20NDIIPP%202012.pdf>.
- Curtis, Bill, William E Hefley, and Sally Miller. *Overview of the people capability maturity model*. Software Engineering Institute, Carnegie Mellon University, 1995. Accessed 1 April 2011.
<http://www.ts.mah.se/utbild/Pv7040/pv7040ht04/mm001.95.pdf>.
- Department of Defense [United States]. *Records Management Application: DoD 5015.02-STD RMA Design Criteria Standard*, 2012. Accessed 27 March 2012.
<http://jitc.fhu.disa.mil/cgi/rma/standards.aspx>.
- Dingwall, Glenn. *Building a digital archives for the City of Vancouver*, 14 September 2011. Accessed 17 September 2012. [http://www.vancouvergis.org/docs/Digital%20Archives%20Preservation%20-%20GIS%20User%20Group%20\(2011-09-14\).PPT](http://www.vancouvergis.org/docs/Digital%20Archives%20Preservation%20-%20GIS%20User%20Group%20(2011-09-14).PPT).
- Dingwall, Glenn, Richard Marciano, Reagan Moore, and Evelyn Peters McLellan. "From data to records: Preserving the Geographic Information System of the City of Vancouver." *Archivaria* 64 (2007): 181-198.
- DLM Forum Foundation. *Moreq2010 - Modular Requirements for Records Systems*, 2011. Accessed 11 June 2011. http://moreq2010.eu/pdf/moreq2010_vol1_v1_1_en.pdf.
- Dollar, Charles, and Lori Ashley. "Digital Preservation Capability Maturity Model." 2011.
- . *A digital preservation maturity model in action*, 2012. Accessed 28 May 2012.
<http://lib.stanford.edu/files/pasig-jan2012/12F2%20Digital%20Preservation%20Capability%20Maturity%20Model%20in%20Action.pdf>.

- Dryden, Jean. "Measuring Trust: Standards for Trusted Digital Repositories." *Journal of Archival Organization* 9, no. 2 (2011): 127-130.
- Duranti, Luciana. "Reliability and Authenticity: The Concepts and Their Implications." *Archivaria* 39 (1995): 5-10.
- Duranti, Luciana, and Heather MacNeil. "The protection of the integrity of electronic records: An overview of the UBC-MAS research project." *Archivaria* 42 (1996): 46-67.
- ECM Maturity Model. *8,500 downloads to date!*, 4 April 2012. Accessed 19 September 2012. <http://ecm3.org/2012/04/04/8500-downloads-to-date-3/>.
- Foscarini, Fiorella, Yunhyong Kim, Christopher A. Lee, Alexander Mehler, Gillian Oliver, and Seamus Ross. "On the notion of Genre in digital preservation." Paper read at Automation in Digital Preservation, July 18-23, 2010, Dagstuhl.
- Galloway, Patricia. "Digital Archiving." In *Encyclopedia of Library and Information Sciences*, edited by Marcia J. Bates and Mary Niles Maack, 1518-1527. Boca Raton, FL: CRC Press, 2009.
- Gartner. (2003). *Magic Quadrant for Integrated Document Management*. Accessed 5 March 2011. http://www.project-consult.net/Files/Gartner_QM2003_IDM.pdf.
- Gillies, Alan. "Information support for general practice in the new NHS." *Health Libraries Review* 17, no. 2 (2000): 91-96.
- International Council on Archives. "Principles and Functional Requirements for Records in Electronic Office Environments: Module 2 - Guidelines and Functional Requirements for Electronic Records Management Systems." International Council on Archives, 2008. Accessed 17 September 2012. <http://www.adri.govt.nz/products/ICA-M2-ERMS.pdf>.
- International Monetary Fund. *Archives of the International Monetary Fund*, 2012. Accessed 2 August 2012. <http://www.imf.org/external/np/arc/eng/archive.htm>.
- InterPARES 2 Project. *Terminology Database*, 2007. Accessed 11 March 2012. http://www.interpares.org/ip2/ip2_terminology_db.cfm.
- Jordan, Paul. *Speech for SAA Chicago*. Society of American Archivists, 2011. http://saa.archivists.org/Scripts/4Disapi.dll/4DCGI/events/eventdetail.html?Action=Events_Detail&Time=226188967&InvID_W=1916.
- Kampffmeyer, Ulrich. *Trends in Record, Document and Enterprise Content Management*. Project Consult, 28 September 2004. Accessed 4 October 2011. http://www.project-consult.net/Files/ECM_Handout_english_SER.pdf.
- . *Enterprise Content Management*, Project Consult. 2006. Accessed 4 October 2011. http://www.project-consult.net/Files/ECM_White%20Paper_kff_2006.pdf.
- Katuu, Shadrack. "Enterprise Content Management— using maturity models to improve the quality of implementation," 2012. [Katuu 2012a]
- . "Enterprise Content Management (ECM) implementation in South Africa." *Records Management Journal* 22, no. 1 (2012): 37-56. [Katuu 2012b]
- Kerzner, Harold. *Strategic planning for project management using a maturity model*. New York: John Wiley & Sons, 2001.
- Kroll, Matthew, David Minor, Bernie Reilly, and Michael Witt. "ISO 16363: Trustworthy Digital Repository Certification in Practice." In *Proceedings of the 7th International Conference on Open Repositories, Edinburgh, Scotland, UK, 2012*.

- Lazorchak, Butch. "Digital Preservation, Digital Curation, Digital Stewardship: What's in (Some) Names?" 2011. Accessed 20 September 2012.
<http://blogs.loc.gov/digitalpreservation/2011/08/digital-preservation-digital-curation-digital-stewardship-what%E2%80%99s-in-some-names/>.
- Lee, Christopher A. "Open Archival Information System (OAIS) Reference Model." In *Encyclopedia of Library and Information Sciences*, edited by Marcia J. Bates and Mary Niles Maack, 4020-4030. Boca Raton, FL: CRC Press, 2009.
- Lee, Christopher A., Richard Marciano, Chien-Yi Hou, and Chirag Shah. "Mainstreaming preservation through slicing and dicing of digital repositories: Investigating alternative service and resource options for ContextMiner using data grid technology." In *iPRES 2009: the Sixth International Conference on Preservation of Digital Objectives*. UC Office of the President: California Digital Library, 2009.
- Lee, Christopher A., and Helen Tibbo. "Where's the archivist in digital curation? Exploring the possibilities through a matrix of knowledge and skills." *Archivaria* 72 (2011): 123-168.
- Liu, Richard Y. L. "Capability maturity model integration: origins and applications." Master's thesis, California State University, Long Beach, 2002.
- MacNeil, Heather. *Trusting Records: Legal, Historical and Diplomatic Perspectives*. Dordrecht: Kluwer Academic Publishers, 2000.
- . "Proving grounds for Trust II: The findings of the Authenticity Task Force of InterPARES." *Archivaria* 54 (2002): 24-58.
- McRoberts, H. A. , and B. C. Sloan. "Financial management capability model." *International Journal on Government Auditing* 25, no. 3 (1998): 8-11.
- Melikhova, Irina. "Long-term preservation (LTP) of digital information: City of Toronto case study." 2010. Accessed 17 September 2012.
<http://www.verney.ca/assets/file/MIPS2010presentations/3A.pdf>.
- MIKE2.0. *ECM Maturity Model (ecm3)*, 2010. Accessed 19 September 2012.
http://mike2.openmethodology.org/wiki/ECM_Maturity_Model_%28ecm3%29.
- Miles, Doug. *State of the ECM Industry 2009*. AIIM, 2009. Accessed 27 September 2012.
<http://www.aiim.org/PDFDocuments/36031.pdf>.
- . *State of the ECM Industry 2010*. AIIM, 2010. Accessed 22 April 2011.
<http://www.aiim.org/Research/Industry-Watch/ECM-State-of-Industry-2010>.
- . *State of the ECM Industry 2011*. AIIM, 2011. Accessed 27 May 2012.
<http://www.aiim.org/Research-and-Publications/Publications/Industry-Watch/State-of-the-ECM-Industry-2011>.
- Moore, C., and R. Markham. "Market leaders emerging in Enterprise Content Management." 2003. Accessed 5 March 2011.
http://www.forrester.com/rb/Research/market_leaders_emerging_in_enterprise_content_management/q/id/30139/t/2.
- Mumma, Courtney C., Glenn Dingwall, and Sue Bigelow. "A first look at the acquisition and appraisal of the 2010 Olympic and Paralympic Winter Games Fonds: or, select *From VANOC_Records As Archives WHERE Value="true";" *Archivaria* 72 (2011): 93-122.
- Murray, A, and M Ward. *Improving project performance using the PRINCE2 maturity model (P2MM)*. Norwich: The Stationary Office, 2007.

- National Archives of Australia. *Guidelines for Implementing the Specification for Electronic Records Management Systems Software*, 2007. Accessed 5 March 2011. <http://www.naa.gov.au/records-management/publications/ERMS-guidelines.aspx>.
- Nguyen L. T., Swatman P. M. C., and B. Fraunholz. *EDMS, ERMS, ECMS or EDRMS: Fighting through the acronyms towards a strategy for effective corporate records management*, 5-7 December 2007. Accessed 25 March 2011. <http://acis2007.usq.edu.au/assets/papers/132.pdf>.
- Pearce-Moses, Richard. *A Glossary of Archival and Records Terminology*. Chicago: Society of American Archivists, 2005.
- Pelz-Sharpe, Alan , Apoorv Durga, David Smigiel, Erik Hartman, Tony Byrne, and Jarrod Gingras. *ECM 3 - ECM maturity model (2nd)*, June 2010. Accessed 5 April 2011. http://ecmmaturity.files.wordpress.com/2009/02/ecm3-v2_0.pdf.
- Peterson, Trudy Huskamp. *Temporary courts, permanent records*. Wilson Center, 2008. Accessed 20 July 2012. http://www.wilsoncenter.org/sites/default/files/TCPR_Peterson_HAPPOP02.pdf.
- Scott-Jones, David. "Implementing EDRM in the Ministry of Defence." *Information Management and Technology Journal* 35, no. 4 (2002): 159-162.
- Seles, Anthea. *Digital records preservation*. National Archives of Thailand, 2012. Accessed 23 September 2012. <http://student.netdesign.ac.th/web553708/images/Anthea.pdf>.
- Shegda, Karen. *Integrated Document Management Software: Perspective*. Gartner, 4 April 2001. Accessed 4 April 2011. <http://www.emory.edu/BUSINESS/et/552fall2001/collaboration/int2.pdf>.
- Shegda, Karen M., Kenneth Chin, Debra Logan, and James Lundy. *Building the Magic Quadrant for ECM 2004*, 12 November 2004. Accessed 5 April 2011. http://www.gartner.com/DisplayDocument?doc_cd=124032.
- Spiro, Linda. *Archival Management Software: A Report for the Council on Library and Information Resources*. Council on Library and Information Resources, 2009. Accessed 23 September 2012. http://www.clir.org/pubs/reports/spiro/spiro_Jan13.pdf.
- Sprehe, J. Timothy. "The positive benefits of electronic records management in the context of enterprise content management." *Government Information Quarterly* 22, no. 2 (2005): 297-303.
- Van Garderen, Peter. *Electronic Records Strategy: Final Report*, 2002. Accessed 24 September 2012. http://siteresources.worldbank.org/EXTARCHIVES/Resources/WB_ERS_Phase_1.pdf.
- . "Archivematica at the City of Vancouver Archives." 2012. Accessed 17 September 2012].
- Ward, Bob. "Enterprise Content Management." *Information Management and Technology* 34, no. 4 (2001): 179-181.
- Waugh, Andrew. "The design and implementation of an ingest function to a digital archive." *D-Lib Magazine* 13, no. 11/12 (2007).
- Wilhelm, Philipp. "An evaluation of MoReq2 in the context of national EDRMS standard developments in the UK and Europe." *Records Management Journal* 19, no. 2 (2009): 117-134.
- Wilkins, Linda, Paula M C Swatman, and Duncan Holt. "Achieved and tangible benefits: lessons learned from a landmark EDRMS implementation." *Records Management Journal* 19, no. 1 (2009): 37-53.
- Yakel, Elizabeth. "Digital Curation." *OCLC Systems and Services* 23, no. 4 (2007): 335-340.

Facilitating the Aggregation of Dispersed Personal Archives

A Proposed Functional, Technical, and Business Model

Christopher J. Prom

Assistant University Archivist and Associate Professor, University of Illinois at Urbana-Champaign

Abstract

We keep records in archives because such institutions are dedicated to preserving authentic evidence of human activity. ‘Cloud’ services pose a direct challenge to the archival mission. Archivists and all of humanity have a direct interest in building tools that help people aggregate, use, and control records they created. This paper outlines the conceptual model one such service, which is dubbed “myKive,” and which is currently undergoing proof-of-concept development at the University of Illinois. After describing its necessity, the paper lists the proposed service’s functions, outlines its core architecture, and describes its development/business framework.

Author

Christopher J. (Chris) Prom is Assistant University Archivist and Associate Professor of Library Administration at the University of Illinois at Urbana-Champaign. He holds a Ph.D. in history from the University of Illinois and also studied at the University of York (United Kingdom). He is a Fellow of the Society of American Archivists and has received several other research fellowships including most recently a 2009-10 Fulbright Distinguished Scholar Award. He maintains the Practical E-Records Blog and an active publication portfolio. His research describes the ways in which archival users seek information relevant to their needs and assesses methods that archivists can use to efficiently meet those needs. He most recently authored a technical watch report for the Digital Preservation Coalition, “Preserving Email.” Chris is also co-director of the Archon™ project, which developed an open source application for managing archival descriptive information and digital objects, and he is a member of the ArchivesSpace project, which is developing a next-generation archival management system. He has served the Society of American Archivists in several capacities. He is currently a member of the editorial board of *The American Archivist*.

I would like to begin my paper with a story. The story demonstrates key challenges faced by archives and archivists in what we might term the cloud era—the era of dispersed digital archives.

Last November, I boarded a train at Union Station in Chicago, Illinois. I had just left a meeting of the Society of American Archivists’ Fundamental Change Working Group. This group was charged with revising the *Fundamentals Series*, which comprises the heart of our society’s publishing program.¹ Everyone at the meeting was acutely aware of two facts: 1) that newly trained archivists need a sophisticated set of digital skills, and 2) that our new instructional manuals must facilitate these skills.

Moving quickly to find a seat on the train, I spotted a person from my University. I’ll call this person “Dr. Important.” After the requisite chic-chat, Dr. Important asked me what I have been working on lately.

“Well, I’ve been writing a guide to email preservation.”

“Oh, that’s interesting. Maybe you can help me.”

¹ The six books that comprise this series are available by visiting the publications pages at the Society of American Archivists website, at <http://saa.archivists.org/store/> (Accessed June 26, 2012).

Who doesn't like to be asked for help? Maybe I could tell Dr. Important how to organize email and export it to a preservation-ready format. If lucky, I might even convince Dr. Important to transfer email to the University Archives, where it would become a public research resource. In this way, it would be accessible much like the handwritten or typescript correspondence from many other important people who had worked or studied at the University of Illinois in past years.

"You see, I went to look for something I sent back in 2009," Dr. Important continued. "I've been keeping a copy of all of my important emails, one folder for each month. But when I went back to find the message I needed, all the folders were gone." Dr. Important told me that technical staff could not restore the emails, which likely went missing during a system migration that had taken place several months prior.

As an archivist, I mourned the death of the evidence and information that Dr. Important had created and cared for over many years. But I felt helpless, and I let the conversation drift to another topic.

This incident, and many others that I could tell from my time at the University of Illinois, illustrate one of the greatest challenges that archivists face: ensuring the preservation of evidence when people's communication tools have, in effect, become their unofficial recordkeeping mechanisms. This problem is particularly pressing because, in most institutions, centralized systems to manage correspondence and other communications are dead or at least have one foot firmly in the grave. Given this fact, what can we (as a profession) do to make sure that usable records are fixed into a medium that will facilitate their perseverance and use?

In order to answer that question, we must understand the ways in which information is dispersed within modern organizations and external social networks. More to the point, we must understand the way in which technology makes information into records that subsist within human social networks. With a better understanding of how records are formed and used within the technologies that facilitate such networks, we will be better positioned to capture and preserve not only information, but also contextual data about how that information was dispersed, used, and reused.

1. 'Record-ness,' Archives, and the Need for a Personal Archives Service

In the cloud era, record capture and preservation systems must take three factors into account: 1) the perceived lack of value accorded to preserving digital communications; 2) the communication and information management practices used by individuals and; 3) the specific ways in which contextual data transforms information into evidence, within human social networks and the technologies that support them. Taken as a whole, the implications of these three factors call into question the continued existence of archives in the cloud environment—if by archives we mean a group of records that are maintained as a collective using the principles of provenance, original order, and collective control.

1.1 The perceived lack of value accorded to preserving digital communications

In 1899, the American sociologist Thorstein Veblen wrote that "the cheap, and therefore indecorous, articles of daily consumption in modern industrial communities are commonly machine products."² Such articles are much used but little valued, at least in a monetary sense. For that reason, they are easily lost or

² Thorstein Veblen, *The Theory of the Leisure Class* (New York: Viking Press, 1967), 161, <http://www.gutenberg.org/ebooks/833>.

discarded. Any American who has eaten at a Fourth of July picnic knows how easy it is to throw dirty plastic utensils and plates into the trash, in spite of their utility when the hot dogs and watermelon were being served.

In post-industrial societies, digital communications comprise one of the cheap, and therefore indecorous, articles of daily consumption. We are familiar with the forms that these materials take: email messages, blog posts, Facebook updates, tweets, online videos. Each can be inexpensively produced with the help of an electronic device. Each is arguably less decorous than the format for communication that it replaced, such as the handwritten letters, illustrated diaries, or professionally produced films in which archives like to traffic.³

Given this fact, one may expect that the greatest challenge in preserving such materials might consist simply in convincing people that their personal digital communications are important enough to preserve. But this is not the case. In the abstract, many people value digital materials highly and keep everything they send or create. However, most of them do not much concern themselves when a system crash sweeps digital records away as in a flood.⁴ This points to an important truism: the broader information ecology in which people work makes it very difficult for both organizations and individuals to identify, capture, and preserve the records that have the most long-term archival value, unless extraordinary actions are taken.

Let me provide a few examples. My own institution, the University of Illinois, formerly made extensive use of college and departmental subject files, documenting faculty teaching, research, service, and administration. I say ‘formerly’ because over the past twenty years these paper-based files have largely disappeared. During the same period, most of our distinguished faculty members stopped keeping systematic correspondence files, aside from messages fortuitously retained within active email accounts. Asking administrators or faculty members to keep records outside of their communication applications (either in paper or in digital form) seems like a fruitless task. First, it would require that the institution implement an expensive software and hardware product, such as an Electronic Records Management (ERM) application. More to the point, implementing such a system would require that people make extensive changes to their work habits and procedures—something that is extremely unlikely in the Facebook Era, with its emphasis on immediate communication and response. Where ERM or ‘document management’ systems have been implemented, we see numerous problems follow. For example, staff in the office of our chief administrative officer (the Chancellor) are worried that email messages documenting critical policy decisions never make their way into the document management system since administrators don’t like to change their work habits and deposit email. Staff members are also worried that the system will not survive the departure of the current records manager.

Underlying this issue, we see a larger problem. While relatively little attention is being paid to building systems to retain substantive records of archival value, our campus has put many resources into building systems that preserve other types of information. For example, most universities maintain expensive business systems that store transactional data relating to financial affairs, personnel management, and student academic records. Most, if not all of the data managed by these systems lack long-term administrative and archival value. The maintenance of these systems is certainly necessary for

³ For one attempt to provide a more decorous platform for personal reminiscence and storytelling in the digital era, see the Cowbird service, founded by Jonathan Harris, at <http://cowbird.com> (Accessed June 26, 2012).

⁴ Catherine C. Marshall, “Challenges and Opportunities for Personal Digital Archiving,” in *I, Digital: Personal Digital Collections in the Digital Era*, ed. Christopher A. Lee (Society of American Archivists, 2011), 99–100.

the daily operation of the University. The data within them is used to produce aggregated management information, which is certainly of long-term archival value.⁵ Similarly, the University of Illinois and many other libraries operate well-designed and successful applications that preserve formal research outputs, such as published scholarly papers. A rich literature has arisen around the development and implementation of these repositories.⁶ By comparison, relatively little has been written concerning the theory and practice of preserving personal or professional correspondence, informal communications, or social media records.

For example, email preservation has not been formally included within the scope of most large-scale digital preservation projects.⁷ As a result, archivists can easily be left in the position of sweeping up the few crumbs of information from a faculty member or administrator's computer or closet, much like I did when cleaning out the office of a distinguished chemist, Stanley Smith, in 2011.⁸ If we want a better future for historical research, we must develop a method for archivists and records creators to work together, so that communications in email systems and social networking technologies can be kept alive long enough to be accessioned to an archives.

1.2 Personal communication and information management practices

An email I received while on sabbatical illustrates how recordkeeping functions have devolved to individual initiative. The author, who wished to remain anonymous, nearly lost his entire email record when his former employer abruptly terminated access.⁹ The prominent American journalist James Fallows relates a similar story concerning his wife's Gmail account, which was hacked, leading to the deletion of its entire contents. Ten years' worth of messages were salvaged only because Fallows took advantage of some personal connections at Google.¹⁰ While these stories are anecdotal, they resonate with our experiences as archivists. Furthermore, they show that efforts to preserve personal records in the era of cloud computing must focus on helping people generate archives that subsist outside of their communication utilities. But what types of services are most needed?

In a two-part article entitled *Rethinking Personal Digital Archiving*, Microsoft Principal Researcher Cathy Marshall notes that most people exhibit an information archiving instinct.¹¹ At an extreme, this tendency can exhibit itself as compulsive information hoarding, either of hard copy or digital materials.¹²

⁵ At the University of Illinois, such datasets are managed by the Division of Management Information. <http://www.dmi.illinois.edu/> Accessed June 7, 2012.

⁶ Charles W. Bailey, "Institutional Repository Bibliography, Version 4," *Digital-scholarship.org*, June 15, 2011, <http://digital-scholarship.org/irb/irb.html>.

⁷ Christopher J. Prom, *Preserving Email*, DPC Technology Watch (Digital Preservation Coalition, December 2011), 4–8, <http://dx.doi.org/10.7207/twr11-01>.

⁸ See <http://www.library.illinois.edu/archives/archon/index.php?p=collections/controlcard&id=10857> for a description of the Smith Papers, and access to some of his digital files.

⁹ Prom, *Preserving Email*, 16.

¹⁰ James Fallows, "Hacked!," *The Atlantic*, November 2011, <http://www.theatlantic.com/magazine/archive/2011/10/hacked/8673/#>.

¹¹ Catherine C. Marshall, "Rethinking Personal Digital Archiving Part 1: Four Challenges from the Field," *D-Lib Magazine* 14, no. 4 (2008), <http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html>; Catherine C. Marshall, "Rethinking Personal Digital Archiving, Part 2: Implications for Services, Applications, Institutions," *D-Lib Magazine* 14, no. 3/4 (March 2008), <http://www.dlib.org/dlib/march08/marshall/03marshall-pt2.html>.

¹² Renae Reinardy, "Information Hoarding: The Need to Know and to Remember," *OCD Newsletter* (Fall 2006): 14–15; Christopher Mims, "The Internet May Encourage 'Information Hoarding' - Technology Review,"

Most people exhibit modest self-archiving behaviors, such as the six methods that Marshall describes in the first part of her article.¹³ Technologies facilitate each of these strategies, as well as other strategies that people have developed since Marshall conducted her research.

However, Marshall finds that archiving technologies or strategies are often used very ineffectively, if the goal of ‘archiving’ is the long-term preservation of evidence of human activity. I have also noticed this problem. One faculty member whom I know contracts with Dropbox to backup her computer desktop, which includes photographs, reports, and working documents. The information on this computer may have some long-term value, particularly if the photographs or documents contain embedded metadata (which is doubtful).¹⁴ But the ‘saving’ is dependent on her remembering to transfer content to the service. Furthermore, the information on her computer is lifeless until it is communicated or shared with someone else. Such communication or sharing takes place via her three email accounts, her blog, and her Facebook page. What this means is that unless the email, blog posts, and status updates from these sources are saved in a fixed format, along with their metadata, the full impact of her work will be lost. Her career and influence will be less well understood than they might be, because the evidential value of her activities will not have been preserved. We will be left with a set of documents that lie mute on an image of her hard drive, lacking the life blood—context—that makes archives a uniquely useful research resource.

Why is saving personal communications important? In the academic realm, I would argue that it is important because libraries in the United States are currently placing a great deal of emphasis on other things. For example, much attention is being given in libraries to the development of data curation and data management services, often with the aim of encouraging faculty to preserve their research data in an institutional or disciplinary repository.¹⁵ Several national programs have been launched, aiming to help librarians become more ‘data-centric.’¹⁶ My institution, for example, launched a “Year of Data Stewardship.”¹⁷ Internationally, a rich literature has developing around this issue.¹⁸

Those practicing data curation should be regarded as valued partners, to whom we can to articulate the archival emphasis on the evidence that makes data useful. The field is open to us. We can develop

Technology Review, September 27, 2010, <http://www.technologyreview.com/view/420947/the-internet-may-encourage-information-hoarding/>.

¹³ 1) System backups; 2) Saving old “My Documents” folders; 3) writing important files to external media such as compact disks or hard drives; 4) emailing themselves important document and attachments; 5) putting materials on social media sites (i.e., Flickr, Facebook); and 6) Saving an image of the entire platform, to be restored in case of system failure.

¹⁴ We are well advised to recall David Bearman’s definition for the word “record” as “communicated information.” David Bearman, “Managing Electronic Mail,” in *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*, ed. David Bearman (Pittsburgh: Archives and Museum Informatics, 1994), 189–90.

¹⁵ Sayeed Choudury, “Data Curation: An Ecological Perspective,” *College and Research Libraries News* 71, no. 4 (April 2010): 194–96.

¹⁶ Council on Library and Information Resources, “Data Curation Initiatives”, 2012, <http://www.clir.org/initiatives-partnerships/data-curation>; Alan Blatecky and Chris Greer, *Data Web Forum Concept Paper*, June 2012, http://www.cni.org/wp-content/uploads/2012/06/DataWebForum_Concept_Paper.pdf.

¹⁷ http://www.cio.illinois.edu/Data_Stewardship. See the Illinois Research Data Initiative blog at <http://blogs.cites.illinois.edu/datasteward/> for more information.

¹⁸ Tony Hey and Anne Trefethen, “The Data Deluge: An e-Science Perspective,” in *Grid Computing* (John Wiley & Sons, Ltd, 2003), 809–824, <http://dx.doi.org/10.1002/0470867167.ch36>; Karen S Baker and Lynn Yarmey, “Data Stewardship: Environmental Data Curation and a Web-of-Repositories,” *International Journal of Digital Curation* 4, no. 2 (2009): [online] and ; Ixchel Faniel and Ann Zimmerman, “Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse,” *International Journal of Digital Curation* 6, no. 1 (2011): [online] provide representative examples of this literature.

complementary tools and services, which will preserve the evidence—the communications—that makes research data interpretable.

Recent archival literature articulates a conceptual framework for personal digital archives. This framework complements the library community's emphasis on data curation. For example, the work of Richard Cox, Kathy Marshall, Cal Lee, and others offers a theoretical basis for preserving the evidence found in dispersed personal digital archives.¹⁹ Similarly, some scholars and members of the popular media have argued that individuals should establish digital legacy plans or even provide instructions regarding the disposition of their digital assets within their estate plans.²⁰

However useful this literature may be, what people most need are practical tools and services that preserve communications and social sharing. By designing and implementing such tools, archivists can work to build trusted relationships with people. If we help people generate true archives from their currently dispersed digital communications, and if we make that archives useful to them during their lifetimes, they (or their heirs) will have the ability and motivation to deposit those archives in one of our repositories.

1.3 How contextual data transform data, information and knowledge into evidence

John Seely Brown and Paul Duguid open their book *The Social Life of Information* with a point that may seem trite: “[I]nformation and individuals are always part of rich social networks.”²¹ However, the implications of this thought are profound for archival practice.²² Elaborating their conception of information's social life, they argue that “documents ... help structure society, enabling social groups to form, develop, and maintain a shared sense of identity.”²³ Brown and Duguid note that information can easily ‘clot’ within formal organizations or records systems, so that critical information is not communicated to those who most need it. (I recently experienced this phenomenon when a key decision-maker never heard about a report the University Archives developed. As a result, he spent over three months duplicating our work.)

On the other hand, data, information and knowledge quickly leak from within organizations to cross-cutting networks of affinity or practice. This is particularly true within an academic setting, but also

¹⁹ Richard J. Cox, *Personal Archives and a New Archival Calling: Readings, Reflections and Ruminations* (Duluth, Minn.: Litwin Books, 2008); Catherine C Marshall, “Rethinking Personal Digital Archiving, Part 2: Implications for Services, Applications, Institutions,” in *I Digital: Personal Collections in the Digital Era*, ed. Christopher A. Lee (Chicago, Illinois: Society of American Archivists, 2011).

²⁰ Evan Carroll and John Romano, *Your Digital Afterlife: When Facebook, Flickr and Twitter Are Your Estate, What's Your Legacy?* (Berkeley CA: New Riders, 2011); Rebecca J. Rosen, “The Government Would Like You to Write a ‘Social Media Will’,” *The Atlantic - Technology Blog*, May 3, 2012, <http://www.theatlantic.com/technology/archive/2012/05/the-government-would-like-you-to-write-a-social-media-will/256700/>.

²¹ John Seely Brown and Paul Duguid, *The Social Life of Information* (Boston: Harvard Business School Press, 2002), xxv.

²² One example of how this insight is transforming archival practice can be seen in the Social Networks and Archival Context Project, led by Daniel Pitti at the University of Virginia. The project seeks to develop a national archival authorities cooperative. Jennifer Howard, “Archive Watch: Building a National Cooperative for Archival Standards - Wired Campus - The Chronicle of Higher Education,” *Wired Campus Blog: The Chronicle of Higher Education*, May 22, 2012, <http://chronicle.com/blogs/wiredcampus/archive-watch-building-a-national-cooperative-for-archival-standards/36368>.

²³ Brown and Duguid, *The Social Life of Information*, 189.

in business and government, since the influential members of an organization typically work within an interwoven professional network. Facebook and other social media applications accelerate the process of information leakage, particularly when an item ‘goes viral.’ But a video need not be viewed by millions of people for information to have a profound impact outside of formal organizational structures. For example, a professor with whom I am currently working exercises a leadership role on numerous faculty committees that exist within our organization, but she also contributes to professional associations, maintains a blog, and helps mentor current and former students via Facebook. For each of the social networks within which she is engaged, she uses particular tools and services, dispersing evidence across multiple systems, and demonstrating her influence in society.

As a profession, we must develop tools and services to capture and preserve something that is more rich than mere data, information, or knowledge. We must find a way to aggregate communicated information that is currently held in dispersed systems, in a way that enhances its value as evidence of people’s activities. As Theo Thomassen noted in his address at the 17th Brazilian Congress on Archival Science,²⁴ the preservation of the evidential value has been one of the archival profession’s defining values. As such, it bears the need for constant reemphasis and renewal.²⁵ At this time, it must be refreshed to address the rise of social networking technologies.²⁶

As Brown and Duguid note, the media that hold information do not merely contain, carry, or convey that information, they also structure its use within communities.²⁷ For example, paper documents are relatively mobile but largely immutable. The 95 Theses of the Augustinian Monk Martin Luther were dispersed through Europe on foot, without much modification, in a fixed format, and at a relatively slow pace. On the other hand, electronic communications are both instantly mobile and highly mutable, particularly when they are shared via email, websites, or, social networking technologies. A blog post can be changed at a moment’s notice, and many versions of it may be distributed or replicated with a single keystroke or computer command.

²⁴ Theo Thomassen, “Archivists and the Private Desire To Be or Not To Be Documented,” *Proceedings of the 17th Brazilian Congress on Archival Science*, (Rio de Janeiro: Associação dos Arquivistas Brasileiros), forthcoming.

²⁵ As Adrian Cunningham has pointed out, the concepts behind of digital preservation and digital curation (such as the Open Archival Information System Reference Model and the Trustworthy Repositories Audit Checklist) provide necessary tools to help ensure that records are authentic. However, they are insufficient to the task of “total archiving” because they do not put sufficient emphasis on pre-custodial interventions that help maintain evidential value. Brian Lavoie, *The Open Archival Information System Reference Model: An Introductory Guide*, DPC Technology Watch (Online Computer Library Center, Inc., and Digital Preservation Coalition, 2004), http://www.dpconline.org/component/docman/doc_download/91-introduction-to-oais; *Trustworthy Repositories Audit & Certification (TRAC) : Criteria and Checklist*, ed. Robin Dale and Bruce Ambacher (Chicago, Illinois: CRL, 2007); Adrian Cunningham, “Digital Curation/Digital Archiving: A View from the National Archives of Australia,” *American Archivist* 71, no. 2 (2008): 530–543; Adrian Cunningham, “Ghosts in the Machine: Towards a Principles-Based Approach to Making and Keeping Digital Personal Records,” in *I, Digital: Personal Digital Collections in the Digital Era* (Chicago, Illinois: Society of American Archivists, 2011), 78–89.

²⁶ The desire for active intervention to shape the record and preserve evidence has been a constant theme since the beginning of electronic records work, but relatively less attention as been paid to the importance of evidence within the context of social networks. For example, Bearman emphasized the evidential value of records within business systems. David Bearman, “Archival Principles and the Electronic Office,” in *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*, ed. David Bearman (Pittsburgh: Archives and Museum Informatics, 1994), 146–75. Reprinted from *Information Handling in Offices and Archives*, ed. Angelika Menne-Haritz (New York: KG Sauer, 1993), 177–93.

²⁷ Brown and Duguid, *The Social Life of Information*, 189–200.

In order to provide useful research materials to future generations, we must find a way to fix these mobile and mutable records into a fixed format, in a defined location, and in a controlled fashion. This should be done in a manner that preserves enough metadata to make the communications understandable within their original social context. Furthermore, archival systems should also store and manage communications sent via multiple systems and networks, so that no one form of communication is given a privileged status. In this way, materials originating from a common source or records creator (and sharing a common provenance) can be treated in a collective fashion—thus preserving them as an archives, here defined as “[m]aterials created or received by a person, family, or organization, public or private, in the conduct of their affairs and preserved because of the enduring value contained in the information they contain or as evidence of the functions and responsibilities of their creator, especially those materials maintained using the principles of provenance, original order, and collective control.”²⁸

2. The Proposed myKive Service

I propose that the objective described above can be accomplished if the archival/memory community develops an extensible self-archiving application. The tool that I have in mind will allow people to aggregate their personal digital communications into a replicated storage location in a fixed, preservation-ready format, where they can control the records. I have given this service the provisional name of myKive (“My Archive”).

2.1 Why a New Service?

Traditionally the ‘fixing’ function has taken place when the papers of individuals are deposited in an archives, likely after sitting in storage for a period of time. The analogue fixing function is method with which archivists are deeply familiar. But whereas every archivist can easily remove records from a closet, relatively few can easily remove them from an email account or a blog. The proposed myKive service will be an open-source software application that archivists can use to help people save their digital communications and other records in a trusted location.

It is tempting to think that people already have access to good tools that allow them to backup and preserve their files. This is simply not the case. Many users write files to hard drives or other locations, but Cathy Marshall’s research shows that they rarely use the software consistently or correctly.²⁹ One faculty member with whom I am working complains that Apple’s ‘Time Machine’ overflowed the backup disk. (My own Time machine recently deleted any backups more than six months old, for a similar reason.) Other people who use such backup devices do not know if they are capturing email, or only desktop documents. In any case, backup programs do not create archives, because materials deleted from the source disk may also be deleted from the backup device, unless the user is technically savvy enough to prevent that from happening.

²⁸ Richard Pearce Moses, *A Glossary of Archival and Records Terminology* (Chicago: The Society of American Archivists Press, 2005), 30.

²⁹ Cathy Marshall, “Ownership, Aggregation and Re-use of Personal Data” (presented at the Personal Digital Archiving 2012, San Francisco, California, February 24, 2012), <http://archive.org/details/personaldigitalarchiving2012pt2>.

Commercial vendors also offer backup services. Carbonite, Mozy, and Crashplan incrementally mirror defined files to an external disk or off-site server. Dropbox can also function as a backup service. While superficially attractive, these services may leave the user extremely vulnerable to data loss or corruption. The end user license agreements (EULAs) for such services typically provide no protection to the user, and the services do not provide the types of digital preservation or migration services that the academic community is ideally placed to provide.³⁰ Similar problems afflict emergent services, such as Backupify and Nuffly, which individuals can use to backup data in cloud-based email or social media.³¹ Therefore, most backup services do not offer a true preservation option.

On the other hand, one might argue that preservation of social media or email is not a problem since the services themselves function as a de facto archives. For example, Facebook's Timeline feature provides users the ability to present a chronological stream of documents, photographs, comments and other materials. This is an example of what Eric Freeman and David Gelertner called a Lifestream.³² Superficially, the features provided by Facebook and Google make these services look like an archives, but as Jason Scott of Archiveteam jokes, "Google is a library or archive like a supermarket is a food museum."³³ While there is much value to the services that social media companies provide, their continuance is predicated upon a business model that exploits personal data as an economic commodity.³⁴ If there is no further monetary value to be gained by maintaining the personal data, one can expect that it will be conveniently deleted or lost. This possibility is highlighted by that fact that in its EULA, Facebook assumes no positive obligation to preserve data. In spite of its size, the entire Facebook service is as susceptible to business failure as any other company. Facebook does, thankfully, provide a method for users to download all of their data, and a current version of the EULA gives people ownership of their records, but few users are likely to know what to do with the data once it has been downloaded.³⁵ For these reasons alone, it would be best if we help people find a way to generate true personal archives.

There is final reason why such a personal archives service is needed. Most people use multiple communication or social media services. By bringing personally created content together in one location

³⁰ For example, under Crashplan the user waives the right to anything but minor remedies and the agreement may be terminated at will by the service provider. Furthermore, the license mandates no positive obligation to provide users access to their own data—not only in the case of business failure, but even during the course of daily business. Code 42 Software, "End User License for Crashplan", July 30, 2011, <http://support.crashplan.com/doku.php/eula>.

³¹ Needleman Rafe, "Backupify Is More Than a Backup Service," *Rafe's Radar / CNET News*, February 3, 2011, http://news.cnet.com/8301-19882_3-20030614-250.html; "Google Gmail Backup Service | Backupify", n.d., <http://www.backupify.com/gmail/>; Nuffly.com, "Services - Nuffly.com", 2011, <http://nuffly.com/services>.

³² Eric Freeman and David Gelernter, "Lifestreams: a Storage Model For Personal Data," *ACM SIGMOD Record* 25, no. 1 (March 1996): 80–86.

³³ Jason Scott, "ArchiveTeam and the Case of the Widespread Recognition," presentation at Personal Digital Archiving 2012. The Internet Archive: San Francisco, CA, February 24, 2012. <http://e-records.chrisprom.com/jason-scott-archive-team/>

³⁴ Facebook, "Statement of Rights and Responsibilities", April 26, 2011, <http://www.facebook.com/legal/terms>; Guilbert Gates, "Facebook Privacy: A Bewildering Tangle of Options," *NYTimes.com*, May 12, 2010, <http://www.nytimes.com/interactive/2010/05/12/business/facebook-privacy.html>; Stevie Marshall, "Commodifying Community: How Contributing to Facebook Is Selling Our Soul and Why We Don't Care," *Online Conference on Networks and Communities*, February 24, 2010, <http://networkconference.netstudies.org/2010/04/commodifying-community/>; James Fallows, "Facebook, Google, and the Future of the Online 'Commons'," *The Atlantic Blogs*, February 3, 2012, <http://www.theatlantic.com/technology/archive/2012/02/facebook-google-and-the-future-of-the-online-commons/252522/>.

³⁵ "Facebook," *Archiveteam*, April 2, 2012, <http://archiveteam.org/index.php?title=Facebook>.

and fixing it into defined formats, we can preserve a more complete picture of a person's life and influence, than if we trust preservation to many service providers, holding data in many locations.

2.2 Functional Description of the myKive Service

In its initial stages, the myKive project seeks to develop an open-source software application that University of Illinois faculty members and students can use to collect email messages, social media, blog postings, reports, desktop files, and other fugitive materials. These records will be saved to a replicated server in an encrypted, standardized, and preservation-ready format. Regular integrity checks will be run to ensure materials are maintained in a trustworthy manner. Content will be stored with sufficient technical and structural metadata to permit its long-term preservation.

Materials in a person's myKive will remain wholly under their control and subject to a strict privacy policy, but stored on a central server. The contents of the myKive account will be replicated to an offsite location. Users will be provided visualization or other tools to make their records useful. People will be provided the opportunity to donate materials to an established archives or manuscript repository, based on mutual agreement and negotiation, at any time they or their heirs wish.³⁶ Once materials are donated to a public institution, they will be managed under an access agreement outlining the terms under which archival users can access the files.

In providing these functions, the myKive project does not seek to replace a person's existing applications or to affect their daily behavior, but simply to provide a method by which they can aggregate and make usefull content that is currently dispersed across multiple services. As part of the myKive pilot project, the University of Illinois will wrap four pieces of software into a web-based dashboard application. Initially, the dashboard will provide access to these tools:

- ***A social media archiving tool*** customized for use by libraries and archives to allow the preservation of account data for multiple users.
- ***An email archiving tool***, which will use the SMTP protocol to transfer all sent and/or received mails (or, optionally, a filtered set) to a designated archival store.
- ***A desktop archiving tool***, which will mirror files from a local computer to a synchronized version control repository.
- ***A communications visulization and mining tool***, which will make materials in the myKive account immediately useful to their creator.

By focusing on three critical formats (social media, email, and desktop files), the project seeks to target format types that are most widely used by people in their personal and work lives. By providing visualization tools, the records will be immediately useful. Preservation is provided as a critical, but secondary, benefit to the end user. While users of the system will surely benefit from its preservation aspects, the project is based on the presupposition that the service will be much more likely to be used if it provides its users with an immediately tangible benefit, rather than a distant, disembodied one.

³⁶ The materials would be managed under a legal agreement/partnership between the records' creator (as a donor), myKive.org (as service provider), and the archives/manuscript library (as beneficiary). More information is available at <http://www.mykive.org>.

2.3 Proposed Technical Model

During the pilot stage, we have developed a provisional technical model. At this time, we anticipate that the application will use and extend existing open source software, which will be wrapped within middleware to be developed by the University of Illinois. This middleware will link the individual system components into an overall ‘archiving’ application, managed from a web-based dashboard. The service core will consist of an application programming interface (API) that links or glues these components together.

Before describing the proposed technical infrastructure in more detail, it is important to note three factors:

1. We propose to build the core application on the LAMP (Linux, Apache, MySQL, PHP) or Ruby on Rails platform, which will ensure the availability of a large community of open-source developers and code contributors.
2. The tool will be utilize a service-oriented architecture and micro-services, which will allow the processing load to be spread among several servers, as necessary.
3. Its interface will include internationalization features (for example, the capability for multi-language support), allowing for world-wide adoption.

Within this overall framework, the application will consist of the following elements, shown in schematic form (see figure 1):

- *Application Core*: user and account management, system security and API. The system will use an MVC framework to segregate application control and data views from the object and data model, which stores and manipulates information in user accounts.
- *Social Media Archiving: Thinkup Application*. The social media archiving component will consist of the ThinkUp application with appropriate extensions and links to our application core. Thinkup provides ways to harvest and aggregate tweets, Facebook content, and Google Plus postings (see figure 2 for a screenshot of a generic Thinkup installation)³⁷ It is being developed by a very active community of developers, led by Gina Trapani of Expert Labs. While the project has not, to my knowledge, garnered any previous use in the archival community, it represents an excellent starting point for social media archiving, provided that such archival activities can be undertaken within a framework that allows for the co-management of records and their eventual donation to an archives. As we get involved with the project, we plan to contribute plugins or other code to the Thinkup Project, and will investigate ways in which Thinkup data can be repurposed and reused alongside other components of the myKive system. Alternately, we will investigate whether Thinkup itself might be extended to serve as the application’s core.
- *Email Archiving: MUSE*: The open-source ‘Muse’ software is suggested as a candidate technology for incorporation into myKive, since it uses an open storage format that facilitates data reuse and transformation, including visualizing, graphing, searching, and browsing.³⁸

³⁷ Expert Labs, “ThinkUp: Social Media Insights Platform”, 2011, <http://thinkupapp.com/>; Mark Sample, “Putting Twitter to Work with ThinkUp,” *ProfHacker - The Chronicle of Higher Education*, October 28, 2010, <http://chronicle.com/blogs/profhacker/putting-twitter-to-work-with-thinkup/28161>.

³⁸ Stanford University, Mobisocial Laboratory, “Muse Home Page”, 2011, <http://mobisocial.stanford.edu/muse/>.

Currently, Muse works via an extension to popular web browsers, such as Chrome, Firefox and Safari, and it installs a JAVA applet on the users computer. A copy of the email is then downloaded to the User's computer, and Muse provides visualization tools to make the email searchable and more useful, as shown in a screenshot from the application (see figure 3). We propose to utilize the MUSE core to build a server-based version of the software. As the project website puts it, "[t]he Muse infrastructure is fairly re-usable and provides support for login, caching, attachments, address book and entity resolution, automatic grouping, text indexing, and other functions."³⁹ If feasible, we will link the Thinkup and Muse data into visualization services at the application core level, providing users a unified method to search, retrieve, view and analyse information across services.

- *Desktop Mirroring Tool:* We propose using standard encryption techniques and secure socket layer technologies to capture and store desktop files from a local computer on a synchronized and replicated server. The SparkleShare application is suggested as a candidate technology (see figure 4).⁴⁰ Sparkleshare is an open source alternative to file sharing and backup applications like Dropbox and Carbonite. In a typically scenario, users of such application download client software and install it on their computers, then upload files to a host machine, perhaps using automatic file synchronization feature. One of the interesting things about Sparklesrare is that the project allows users to upload files to any host machine using git, provided that the host machine is configured as a git repository.⁴¹ We propose that by establishing myKive as git repository, then integrating the Sparkleshare synchronization tools into our myKive application core/dashboard, we can add automated desktop file backup, synchronization, and even version control, into the system. In theory, users will be able to choose whether to keep a complete record of the desktop over time, or only the latest version of the files.
- *Dashboard with Visualization Tools:* Users will be provided a dashboard application to manage the system components described above. Once they complete initial setup, users will need to perform little or no maintenance, other than keeping passwords current. However, we believe that be providing visualization and data mining tools, such as those integrated into MUSE and Thinkup, we will provide users a reason to remain interested in and use the application. Over time, we propose to integrate other data mining and visualization tools into the service. This work is anticipated to take place after the pilot service has been developed.

Finally, the core API will include a plug-in architecture, similar to that used in wordpress or other web applications, so that other types of material can be preserved via the service. This will allow people from the computer science, archives, library, and digital curation communities to extend the application. Using those extensions and plug- ins, people will be able harvest records from other services into their myKive.

³⁹ Muse help pages, <http://pepperjack.stanford.edu:59992/muse/help>

⁴⁰ See the Sparkleshare project website at <http://sparkleshare.org/>; See also Danny Steiban, "Sparkleshare – A Great Open Source Alternative To Dropbox." Make Use Of Blog: July 7, 2011. Available: <http://www.makeuseof.com/tag/sparkleshare-great-open-source-alternative-dropbox-linux-mac/>

⁴¹ See [http://en.wikipedia.org/wiki/Git_\(software\)](http://en.wikipedia.org/wiki/Git_(software)). Git is an open source revision control system, typically used for source code management in open source projects. However, it can be used as part of any distributed file sharing project, and has the advantage of including automatic revision controls, which can be used in myKive to keep a record of—and copies of—changed or deleted files.

For example, the following record types might be included in an individual's myKive, once appropriate extensions are developed:

- Blogs, using backup tools such as ArchivePress and WordPressDatabase Backup;
- Photographs, using capture tools such as parallel-flickr;
- Web pages, using harvesting tools such as wget, warc, and NutchWax; and
- Personal reference/citation libraries, using tools such as Zotero's application programming interface.

2.4 Business Model and Sustainability

At this time, myKive is a pilot service, in the very early stages of its development at the University of Illinois. If the pilot is successful, we anticipate that the project will garner widespread interest, and may serve as the basis for a collaborative international project.

Since the project will use or extend existing open-source projects and tools, it will itself be made freely available via github, a code repository and version control system. This will encourage a collaborative development process.

Based on the results from this project, we will consider launching a not-for profit community resource via the www.myKive.org website, in conjunction with a non-profit partner such as Duracloud/Duraspace and/or the Internet Archive.⁴² In this way, the service will be useful not only at the University of Illinois, not only in the State of Illinois, not only in the United States of America, but hopefully, worldwide. The proposed myKive service would include the following elements:

- Competitive pricing model vis-à-vis commercial backup services,
- Encrypted data transfer and storage, and
- Integrated preservation management features such as automated checksum generation and monitoring.

Initially, subscribers would self-register and pay a monthly fee, or their parent institution would pay on their behalf. For people or institutions who subscribe to the service, any records in their myKive account will be stored in an encrypted format and will remain wholly under their control, subject to a strict privacy policy. The service aims to be self-funding once initial research and development have been completed.

I can also foresee a long-term support model. Optionally, users will be able to donate materials to an established archives or manuscript repository, based on mutual agreement and negotiation, at any time they wish. The materials will then be managed under a legal agreement/partnership between the records' creator (as a donor), www.myKive.org (as service provider), and the archives/manuscript library (as donee). Once materials are donated to a public institution, they will be managed under a deed of gift and access agreement outlining the terms under which members of the public can access the files, following standard archival practices.

⁴² Duraspace Foundation, "Duracloud Website", 2012, <http://www.duracloud.org/>.

3. Conclusion

In 1965, the founding archivist of the University of Illinois Archives, Maynard Brichford, wrote that: “Wherever the archivist may be located organizationally, he should be out of his office two-thirds of the time. While processing must be done in the Archives, the archivist should define and standardize processing procedures so that he may spend his time in locating the historical documentation relating to the activities of the university’s staff and students. Effective appraisal must be done in offices, storerooms, stockrooms, and basements.”⁴³

This advice is no less valuable today than it was in 1965.

The myKive service seeks, ultimately, to provide people the ability to control their own records. But it will also provide the archival profession the ability to spend a good portion of our time ‘outside the office’ and to help people save records that are worth saving, in a way that rescues them from dispersion, collecting them into a real archives that is based on provenance, original order, and collective control. We will still rescue materials from storerooms, closets, and basements, but we will also help people rescue materials from Facebook profiles, email listservs, and blog commenting systems. With a service like myKive, we can help empower individuals to save these materials in a preservation-ready and accessible format, where they will be more useful to them today. And in the future, we can empower them donate those materials to an archives or manuscript repository—at a time of their choosing. If this idea excites you, I invite you to support the project and to lend your own perspective and expertise as it develops over the next year, as an initial project partner.

⁴³ Maynard Brichford, “Appraisal and Processing,” in *University Archives*, Allerton Park Institute Proceedings 11 (University of Illinois Graduate School of Library Science, 1965), 46, <http://www.ideals.illinois.edu/handle/2142/433>.

Digital Objects as Forensic Evidence

Constructing and Evaluating Digital Evidence for Processes

Carsten Rudolph and Nicolai Kuntze

Fraunhofer Institute for Secure Information Technology SIT

Abstract

Digital evidence cannot be restricted to single evidence records. A large part of forensic investigations is concerned with relating events to each other and relating events to persons that have initiated them. This paper proposes different options to use meta-data models and meta-data graphs to support this task and also discusses the advantages and disadvantages of the different approaches.

Author

Dr. Carsten Rudolph received his Ph.D. in Computer Science at Queensland University of Technology, Brisbane in 2001. Since then, he is working at the Fraunhofer Institute for Secure Information Technology SIT where he is now the head of the research department on Secure Engineering. His research concentrates on information security, formal methods, security engineering and cryptographic protocols with a strong focus on hardware-based security and digital evidence. Currently, he co-ordinates the EU FP7 project SecFutur on security engineering for embedded systems and he is involved in various other international and German research initiatives.

1. Introduction

The task of reliably documenting processes in technological systems is complex and is usually executed separately on different levels. That is, network events are collected and evaluated independently from information on application level processes. Existing logging can provide audit trails for subsequent evaluation of what has happened. Nevertheless, even when considering only basic requirements on forensic readiness, such audit trails can usually not be considered to be sufficiently secured against manipulations in particular during collection of the information. However, recent work on creating secure forensic evidence shows how single data records can be produced and stored in a secure way and how digital traces of evidence can be constructed.

This paper proposes a technological basis to construct digital evidence for several linked events in order to reliably document a complete process. The linking of events can either occur through information available in the events themselves or by indirectly linking events through meta-information that again needs to be recorded by another securely documented event.

One example for such a chain of events is the process leading to measurements done by calibrated devices. It is obvious that data records produced by such a device need to be protected in order to be suitable as digital evidence. However, the event of measuring also needs to be linked to the process of constructing and calibrating the device.

Another example is the documentation of processes in enterprise networks. Reliable meta-data information can link actions on a network level (e.g., network access, user login via a central authentication server, etc.) to application level events executed by the same user using a particular device. Current logging systems can provide forensic data for some events on the network level. However, the linking to application processes requires additional information that is not available in current enterprise networks.

The approach discussed in this paper combines hardware-based security solutions with so-called meta-data access protocols and meta-data graphs visualizing the information linking of events. This visualization can also support the evaluation of the collected evidence by clearly showing the relations between the different events.

The use of meta-data information provides a much broader view on what kind of information given by digital evidence. Furthermore, it shows the large scope of associations that can be expressed by digital evidence and that therefore needs to be considered in the evolution of digital forensics and also in the view of digital data by lawyers and courts.

2. Two Examples for digital processes

This section discusses two scenarios for digital processes. In both scenarios there are obvious requirements for documenting essential parts of the process. Furthermore, it is not straightforward to relate events to a particular instance of the process and to relate activities by a particular person to the instance of a process. Usual logging activities might not be sufficient for a forensic re-construction of instances of the processes. In general, it is not obvious how events can be linked. One task of digital forensics is concerned with linking events and relating technical events to each other and to physical events. In particular, in the case of conflicts, relating technical events and digital documentation to actions executed by persons can become relevant.

The two scenarios show that different characteristics of a digital process can be relevant. The first process occurs in an enterprise network with remote access. All physical human activities in the process are concerned with people accessing a computer connected to the network. The second example also includes other technical activities like the installation of devices and the configuration and calibration of these devices.

2.1 Nested authentication in enterprise networks

This scenario assumes that a database application is installed on a server somewhere in a protected part of an enterprise network. Further, it is assumed that transactions on the data stored in the database can be critical. Examples for such critical actions can be access to confidential information, changes to accounting data, financial transactions, changes to banking and payment information, but also technical actions in industrial automation scenarios. The actual type of transaction is not relevant for the discussion on digital processes within this example. The actions discussed here are concerned with user authentication and access to particular computers on the network. However, it should be noted that transactions within one application can also be seen as a digital process where similar requirements can occur on the application level. The main goal of the digital evidence should be to prove which individual was responsible for invoking a particular transaction.

The example scenario consists of the following possible sequence of events. All human actors can be employees of a company. This is not about escalation of privileges or other typical security attacks. All actors can own the right to execute all transactions in the process. The goal is to be able to relate transactions to persons:

1. Actor Alice remotely logs in to the enterprise network using a virtual private network VPN. She uses her own credential to securely establish the VPN connection. The credential can be implemented as a password, as a digital credential stored on a SmartCard, or some combination

of password with a physical token. After authentication, she has access to a protected part of the network behind a firewall, e.g., she can have access to a particular computer in the internal network.

2. In order to access the critical application and execute a transaction Alice establishes a remote desktop application to a server in the protected network in order to get access to the graphical user interface of the application. The server initiates a second authentication process, e.g., by requesting a pair of user-name and password. Now, Alice can either use her own credentials or another user-name and password known to her for some reason (e.g., by her colleague Bob).
3. On the server, Alice can now start the application. Once more, authentication might be requested, this time by the critical application itself. Again, Alice can either use her own credentials or another user-name and password known to her.
4. Finally, Alice executes a critical transaction and terminates the connection.

All steps can be logged and we assume for the discussion that secure logging mechanisms exist and are used. In the case of a conflict (or simply for documentation) it can be necessary to know who actually initiated the critical action using the application in question. The documented events show that someone was logged in to the application and the transaction can of course be related to the person whose credentials were used to log in to the application. Now assume that Bob's credentials were used and Bob denies having initiated this transaction. Various attack vectors are available to Alice to steal a password from Bob. Examples include key-loggers installed on Bob's computer or simply spying on Bob, which is perfectly possible if Alice is a co-worker. Also social engineering techniques can be applied. Examples for such techniques include phone calls apparently coming from a technical administrator asking for a password.

Clearly, digital forensics needs to be used to resolve this conflict. In principle, all logging information to determine the actual sequence of events should be available on the system. However, relating these different events and showing the correct sequence is not straightforward and it is not obvious that the logged information is actually suitable to correctly relate events. One can however expect that events do contain some information that can relate events either directly (e.g., by comparing IP addresses, user-names, process IDs, etc.) or indirectly (e.g., via events logged by some authentication server or directory). Nevertheless, even though this information can be available, actually identifying the correct parameters and re-constructing the process is difficult. Furthermore, relations between events are seldom unique and just by evaluating events in the context of some static information on the enterprise network can result in ambiguous results.

Clearly, the re-construction of a digital process can be supported by meta-information (or meta-data) for the different events. This meta-data should identify those parameters that can be used to relate events and describe the relations that can occur. Parts of this metadata can be available on so-called configuration management databases (CMDBs). However, the information in CMDBs is concentrated on the infrastructure and does not include meta-data for applications and other higher layers. Thus, additional meta-data can be necessary and this meta-data will most probably evolve faster than the information on the network infrastructure. Thus, no static meta-data set can be assumed and in addition to the logs also a currently valid meta-data set needs to be stored.

In addition to the construction, storage and evolution of meta-data sets, this information can also be used to analyse possible evidence on processes. Knowledge of the meta-data showing relations between

events can be used to determine (at the time of developing the system/network) if available meta-data will be sufficient to provide all relevant relations between logged events.

The complexity of the forensic evaluation in this example should not be underestimated. Many other persons could have been logged in at the same time and even Bob might have used the critical application himself in parallel to the use by Alice impersonating Bob. Furthermore, once the actual process leading to the critical transaction has been identified, it is still not straightforward to conclude which action has been involved by which person. In addition to the identification of the correct sequence of events in the digital process, it is necessary to evaluate which conclusions can be drawn. This process can include a risk analysis for the different actions. One example can be to distinguish different security levels for the authentication. Obviously, digital evidence for processes can only be one element in the dispute resolution.

Another property of this example is, that relevant events occur in a rather short time-frame, restricted by the duration of authenticated sessions in all different parts of the network (VPN, remote desktop, application). The example in the following section is different as relevant events can be spread over a much longer time.

2.2 Digital speed cameras

This example revises a scenario previously introduced by the authors. The scenario consists of a digital camera to record speeding and process picture data, a server collecting data records with the computed information on speed, number plate etc., network components for communication, long-term storage for data records, and finally, evaluation components.

The existing technical solution provides secure digital evidence bound to the status of the device. Furthermore, a framework exists that enables the control of the validity of the chain of evidence at run-time. However, in reality this system is integrated in a much more complex infrastructure and the reported valid chain of evidence does not provide sufficient information for all situations. Logging data and monitoring can provide information that is not directly related to the actual chain of evidence showing the validity of the measurement, the picture and the evaluation of the values. However, in addition to the actual chain of evidence it can be necessary to also consider supporting processes and document these processes. Such processes include maintenance of the camera, security management for the server and the infrastructure the server is running in, access control (technical, but also physical), and administrative processes like issuing speeding tickets and factorization.

Clearly, not all of the processes mentioned above are relevant for forensic evaluation and not all of them need to be related to the original chain of evidence. However, depending on the character of a dispute some events need to be considered and also related to each other and to particular instances of the main chain of evidence. Again, it might not be possible to maintain a meta-data graph at run-time for all the different events and store the complete history of meta-data. Therefore, different ways need to be found to store meta-data information such that they can be used to evaluate stored log data.

In addition to the storage of meta-data sets to enable off-line calculation of the relations between events, also run-time aspects can be relevant. In particular in this scenario of speed cameras it can be relevant to continuously check that the collected information constitutes valid chains of evidence. Checking the validity of digital signatures on single data records is one part of the validation. However, as described above, all relevant parts of the process should be considered. Thus, a meta-data graph can be constructed at run-time that shows the relations between different events. Evaluation of this meta-data

graph can show that collected evidence is sufficient when seen in the context of the current meta-data specification.

3. Digital evidence for processes using metadata information

This section first revises existing technical solutions and then discusses two different ways to deal with metadata information to construct and evaluate digital evidence for processes.

3.1 Technical building blocks

3.1.1 Trusted computing and digital evidence

One important aspect of the generation of digital evidence is the status of the device used in the process. The software and configuration used to produce evidence needs to be presented and linked to the individual record. One simple scheme hereby is to include software name and version number as a simple string of text in each evidence record. This first (and often used) approach allows for uncertainties with respect to updates and various attacks on the evidence records. Just naming the software is not sufficient if the device can be manipulated. Stronger means of protection are therefore required to reliably document the software and configuration of the particular evidence generator. To provide proof on the actual state of the evidence generator, trustworthy reporting in the device is required. The Trusted Platform Module that is standardized by the Trusted Computing Group TCG introduces a core root of trust for measurement that establishes the foundation to report on the status by creating a chain of trust. This chain of trust can be reported to external entities to allow for a verification of the evidence generator. This verification process is called Remote Attestation.

Application of remote attestation allows for a session-based or per-record scheme to protect digital evidence. The *session-based* approach relies on an initial attestation of the system and a session bound to the individual evidence generator and status. Each evidence record is then cryptographically bound to this session and therefore to a particular system state. The second *per record* scheme involves an attestation process for each evidence record. As in the basic remote attestation, an external random number generator is involved, and longer delays as well as higher bandwidth utilization are to be expected. More advanced schemes allowing for scalable attestation schemes can be applied.

The approach to hardware-based evidence generation linking the evidence to the platform state was first presented in SADFE 2010.

3.1.2 Metadata and IF-MAP for the construction of digital chains of evidence

A technical infrastructure for a secure collection of events and for storing them with matching meta-data was shown in SADFE 2011. The interface to meta-data access points (IF MAP) as an established industry standard and part of the Trusted Network Connect (TNC) protocol stack defines and supports event distribution and correlation in the domain of network access control but can easily be extended to support other types of events and create event graphs representing relations between different types of events. To use IF-MAP to relate events and correlate pieces of evidence was also first proposed in SADFE2011. By now, IF-MAP meta-data models have been extended to include events on mobile devices and a first draft of a meta-data model for industrial control systems also exists. For the creation of digital evidence, these meta-data models need to be extended for each scenario. In particular, including information for events

provided by particular software or domain-specific meta-data models for particular application domains can provide cross-layer relation of events.

3.1.3 Run-time evaluation of metadata versus offline computation of meta-data graphs

The precondition for all evaluation of meta-data is the existence of a meta-data model that describes meta-information for events and thus expresses relations between different events. In the IF-MAP standard, the meta-data model is described in an XML following a standardized XML schema. Relations between different entities in the meta-data model can be shown as a meta-data graph. The complete meta-data model defines the biggest meta-data graph for a system. In reality, this biggest graph will never occur, but all meta-data graphs representing a concrete system state are sub-graphs of this biggest graph.

3.1.4 Computation of metadata at run-time and storage of sub-graphs

The run-time evaluation of metadata can have different goals. First, meta-data graphs can be used to validate digital evidence for processes at runtime and ensure that the stored evidence together with the meta-data information can indeed be considered valid evidence. Second, an evaluation component could subscribe to relevant events in the graph and store each change in the graph together with collected digital evidence records. Then, these smaller meta-data graphs can directly be used to re-construct digital processes from events and relate events of the process or the complete process to persons responsible for initiating the events.

The computation of meta-data at run-time has two advantages. First, it is possible to continuously check that collected digital evidence is not corrupted and can in case of a dispute indeed be used to re-construct processes. Second, forensic evaluation is much easier if relevant subsets of the meta-data graph are stored together with the actual evidence. It can also be used to automate the evaluation of digital evidence and to generate enriched meta-data graphs that provide direct access to the relevant data records or log entries.

However, the run-time monitoring approach also has restrictions. Additional reporting of logged events needs to be done, as the server building the meta-data graph needs to be informed about all events that are logged. This requires additional technical implementations. Even worse, privacy regulations might require to keep different logging information strictly separated and only allows forensic investigations to relate different digital evidence records. In addition, the amount of meta-data to be stored in addition to actual evidence records can get very high. Experiments have shown that already a meta-data graph representing services running on one single smart-phone consists of several hundred nodes and edges. Furthermore, such meta-data graphs can be very dynamic. To support evaluation of digital evidence in the context of the current meta-data graph, all changes need to be stored. In larger enterprise networks, meta-data information would considerably increase the amount of logging data to be stored.

3.1.5 Off-line evaluation of metadata

In contrast to the scheme described in the previous paragraph, in this scheme only the meta-data model is stored and not the dynamic meta-data graph. The meta-data model is also subject to changes, for example whenever the network infrastructure is changed, new applications are installed, or new organisational roles introduced. Off-line evaluation means that event logs are collected and stored together with the

currently valid meta-data specifications. Each change in the meta-data specification needs to generate a revision of the stored meta-data specification and this change needs to be logged in a way that there for each entry in the event log files the correct meta-data specification can be identified. One possibility is to include in all logs *change meta-data* events with unique identifiers for the old and new meta-data specifications. However, these changes occur at a very low frequency when compared with the meta-data graph. Thus, storing all changes in the meta-data model can easily be stored together with the remaining digital evidence in a forensic database.

The consequence that results from the more efficient monitoring is a more complex evaluation process. Evaluating digital evidence in the context of a meta-data model means to construct a meta-data graph from the log information. In principle, this can be achieved by replaying or simulating all the logged events and constructing the metadata graph for this behaviour. In contrast to the run-time construction of the meta-data graph, relevant log events cannot be chosen beforehand and all events need to be computed to explore the metadata graph and then to re-construct the process leading to the particular events under dispute.

3.1.6 Mixed scheme: Meta-data computation at run-time with restricted storage of meta-data information

In the case that critical processes can be identified on the level of meta-data information, it is possible to construct a scheme that uses run-time computation of the meta-data graph in combination with off-line evaluation. The complete meta-data graph is computed at run-time (e.g., by using an IF-MAP server) but the graph is not stored. Instead, it is continuously evaluated using rules that identify potentially critical processes. Then, the logged events leading to this particular meta-data subgraph are identified and tagged as events belonging to one instance of this process. It should be noted that single events can belong to several such processes. Then, in the case of a forensic investigation, the stored meta-data model can be used for a targeted reconstruction of the meta-data graph for this process. This subgraph then shows how the log events are related without the need of a long-term storage of large meta-data graphs.

4. Conclusions

Standard components can support the development of systems that provide secure digital evidence and also can support digital forensics by relating events. If devices, software and applications support the generation and distribution of event information in a standardized format (like IF-MAP) all components of the framework do not interfere with the infrastructure. Standardized components exist and MAP clients reporting events in the standardized format are developed for PCs, network components (routers, switches), mobile systems, and also for components of industrial control systems. In the long run, this development enables the consideration of digital evidence for critical processes already at design time and in addition can support forensic investigations

Forensic Barriers

Legal Implications of Storing and Processing Information in the Cloud

Aaron Alva, Scott David and Barbara Endicott-Popovsky

Abstract

Litigation represents an out-of-the-ordinary event for organizations. A significant percentage of the costs of litigation are associated with the pre-trial processes through which facts are gathered by the parties to be presented to the judge or jury. As organizations become increasingly dependent on cloud services, they will require increasing levels of assurance that cloud services will be performed in accordance with business, legal and technical standards sufficient to enable preparation and support for litigation. This paper examines some of these issues, suggesting a framework for thinking about this problem, suggesting avenues of future exploration.

Authors

Aaron Alva is a candidate of the Master's of Science in Information Management at the University of Washington. He will be attending the University of Washington School of Law next year to begin a joint-Juris Doctorate. He earned his Bachelor's at the University of Central Florida studying political science with a minor in digital forensics. His interests are in cybersecurity law and policy creation, particularly digital evidence admissibility in U.S. courts. Aaron is a recipient of the National Science Foundation Federal Cyber Service: Scholarship For Service, and is a member of the American Bar Association Information Security Committee.

Scott David, J.D. is the director of the Law, Technology & Arts Group at the University of Washington School of Law. Scott's practice focuses on intellectual property and technology counseling and transactions, and he has become a leading expert in these areas nationally and internationally. He is currently developing programs at the intersection of law, technology and arts, including those affecting traditional IP areas (copyright, patent and trademark) and emerging areas involving data, information and identity issues. Scott has been a partner at K&L Gates since 1992. Prior to joining K&L Gates, he was an associate at Simpson Thacher & Bartlett in New York. He has been active in the World Economic Forum (Davos), Identity Commons, ABA Cyberspace Committee, and Workboat (technology/arts organization) among other activities. Scott received his Juris Doctor from Georgetown University (1986), and his LL.M at New York University (1990).

Dr. Barbara Endicott-Popovsky is Director for the Center of Information Assurance and Cybersecurity at the University of Washington, designated by the NSA as a Center for Academic Excellence in Information Assurance Education and Research, Academic Director for the Master's in Infrastructure Planning and Management in the Urban Planning Department of the School of Built Environments, holds an appointment as Research Associate Professor with the Information School, and was named Department Fellow at Aberyswyth University Wales (2012). Her academic career follows a 20-year career in industry marked by executive and consulting positions in IT architecture and project management. Her research interests include enterprise-wide information systems security and compliance management, forensic-ready networks, the science of digital forensics and secure coding practices.

1. Introduction

For organizations, litigation represents an out-of-the-ordinary event. Unusual events typically attract little attention and few resources, with focus instead being placed on the normal operating demands of running an organization. As such, litigation is not often adequately planned. Day-to-day issues and concerns frequently take precedence. Because litigation is unusual, it is frequently costly. Some of that expense can be mitigated with better pre-planning. Fortunately, many of the techniques and approaches that can be implemented to control future litigation costs can also yield day-to-day benefits that arise from better “data system hygiene.” These “dual use” techniques and approaches can therefore be better justified as having value even in the absence of future litigation.

A significant percentage of the costs of litigation are associated with the pre-trial processes through which facts are gathered by the parties to be presented to the judge or jury. These “facts” are discerned from information provided by parties to a lawsuit and other non-litigating parties that are called upon by the court to provide information for use by the parties in the lawsuit. Thus, discovery costs may be borne by parties whether or not they are directly involved in a lawsuit.

In either event, the “facts” are conveyed to the court through the discovery process where information in various forms (such as written materials, testimony, recordings, physical evidence, etc.) is requested for use in the proceedings. Those systems that are able to produce information most readily, upon demand, will be best able to respond to such discovery requests. Not surprisingly, they will also be those that are frequently the most comprehensively managed, which yields myriad day-to-day benefits to businesses. These are also areas in which litigation planning and good “data hygiene” can merge. For this reason, the term “forensically ready” was coined.

Notwithstanding the potential benefits both within and outside of the litigation context, only 56% of organizations polled have, or are in the process of developing, a defined information retention policy.¹ For organizations storing information in the cloud, only 16% stated an Electronic Discovery, or “eDiscovery,” plan was in place before moving data to the cloud.² Additionally, one study has noted that a large majority of cloud providers, and general counsel and business managers from organizations, did not understand general eDiscovery requirements.³ These numbers suggest that there are broad possible opportunities for improvement in the area of discovery, and in the management of cloud-related risks more generally.

In an effort to identify areas of possible improvement in the ways that organizations handle data, and toward the greater goal of reducing organizational risks and costs associated with data and information handling, this paper will consider selected examples of such risks, using discovery requirements as a source of objective requirements from which cloud-related planning can be prompted.

Among the risks that migration of data storage and processes to the cloud presents are new functional, operational, legal and administrative barriers that affect the collection and processing of electronically stored information, or “ESI” which resides in the cloud. These barriers in turn can result in new legal risks, particularly where there is insufficient anticipation of how the barriers might impede

¹ Symantec. *Information Retention and eDiscovery Survey Global Findings 2011*. Accessed June 27, 2012. https://www4.symantec.com/mktginfo/whitepaper/InfoRetention_eDiscovery_Survey_Report_cta54646.pdf.

² Barry Murphy, “e-Discovery in the Cloud is Not As Simple As You Think,” *Forbes*, November 29, 2011, accessed June 14, 2012, <http://www.forbes.com/sites/jasonvelasco/2011/11/29/e-discovery-in-the-cloud-not-as-simple-as-you-think/>.

³ “Results of the 2012 eDSG Investigation of Cloud Service Providers and eDiscovery,” last modified March 7, 2012. <http://ediscoveryconsulting.blogspot.com/2012/03/results-of-2012-edsg-investigation-of.html>.

responsiveness to compelled requests for information. These legal risks can arise regardless of whether or not such data is compelled for a court case, since there are several related scenarios under which an organization might be called upon to provide information (including various administrative proceedings, regulatory review and investigation, contractually-defined performance audits, etc.).

This article will explore some of the risks involved with processing and storing data in the cloud, particularly for Electronically Stored Information, or “ESI” subject to electronic discovery. We will 1) discuss legal instruments that control cloud provider/consumer relationships; 2) detail barriers that increase the legal risks of storing information in the cloud, including implications of U.S.-based eDiscovery requirements; and 3) explore the balance between the potential costs of a purely “reactive strategy” resulting from a lack of forensic readiness and the costs incurred with respect to a “planned strategy” by putting a forensic- and discovery- ready system in place.

2. Legal Control Structures

Cloud services are used by organizations as part of their normal operations. Cloud services agreements are typically entered into “online” without the formalities (and generally without the negotiation) that might characterize more traditional service agreements that document “outsourced” computer services. Whether entered into online or through more traditional means, the agreements document and memorialize the respective rights and duties of the cloud provider and the party using the service.

Most of these agreements purport to cover cloud provider obligations in the context of litigation and other compulsory processes; however, because they are prepared by the same service providers that offer their cloud services, the terms typically emphasize the interests of the service provider. In other words, they are typically drafted to relieve the service provider of obligation or liability to the extent possible. This is not surprising, as all parties prepare their contract “offer” in a manner that is to their advantage. The issue is whether any terms can be modified, and if they cannot, whether the service provider’s service offering is sufficiently attractive that the risks to the service customer from any non-negotiable, or un-negotiated terms remains acceptable to the customer.

In any event, it is clear that prior to any actual legal action, there are agreements that control the duties and rights within the relationship between the cloud provider and customer. Those agreements will affect the organization’s rights, duties and obligations both during its ordinary operations and in the event that the organization is involved in legal action. This paper advocates that both settings should be taken into consideration in evaluating cloud service agreements. It also posits that organizations that are not in a position to “negotiate” cloud service terms can still take unilateral actions to protect their interests such as structuring their data operations and their contracts with their other suppliers and customers to address issues resulting from the standard cloud service rules.

The use of a third party service to store the data of an organization introduces a host of new issues that should be considered before migration to the cloud occurs, and with each successive step through which organizational dependence on cloud services increases. These issues should inform the specific arrangements through which cloud services are procured, even where a cloud customer does not have the opportunity to negotiate terms. Where possible, the issues should be addressed in the agreements through which the cloud services are received. Where those contracts cannot be negotiated, consideration should be given to limiting the manner in which the cloud services are used to protect the company against the risk of using a service that might not deliver the needed services in a given context that is important to the organization. The ability to respond to requests for information in the context of legal processes may be

one of these contexts, particularly for organizations that operate in regulated industries or sectors where litigation is more common.

Ensuring that cloud services used are forensically ready to the satisfaction of the organization—whether by the promises of the cloud provider made in its agreement or by the unilateral internal action of the organization itself through unilateral action to protect its interests while using a service—would help to mitigate the risk of the inability to produce forensic evidence.⁴

3. Service Level Agreements

“Service level agreement” (SLA) is the term applied to those contracts that document the undertakings and control the relationship between the customer and the provider of a service, including those various services provided by cloud service providers. The terms agreed to within the SLA should typically provide guidance on how a production request for evidence will be handled; with the typical provision usually crafted to relieve the service provider of additional burdens to the extent possible. This approach is, unsurprisingly, taken by cloud providers to help control cloud provider costs and liability, which would otherwise affect the price of cloud services. It is also done in order to enable the service to achieve the benefits of scale; an ability that might be diminished if the cloud provider sought to address the various details of reporting obligations in myriad administrative processes and judicial courts in multiple jurisdictions.

A large majority of cloud forensics survey participants noted that tools, techniques and other information for forensics investigations should be included in SLAs.⁵ It is important and relevant to an organization’s business decisions to provide the decision makers involved in strategy and contract administration with an overview on the particular legal implications of SLAs and how SLAs affect the relationship between the customer and provider in both ordinary and out-of-the-ordinary situations. Those legal provisions have business and other economic and strategic implications for the organization.

SLAs can be of importance particularly when setting terms for collection of forensic data or for eDiscovery. SLAs typically delineate obligations of the cloud provider with respect to uptime for the customer, and other relevant commercial terms.

To the extent that they are covered at all, provisions dealing with litigation-related data requests are often given limited treatment, sometimes embedded within a general statement that the company will comply with all laws and legal process. Some commenters have suggested that a more specific treatment of the topic is appropriate, and several have advocated that agreements should include language on how evidentiary records requests will be handled,⁶ including the processes for conducting investigations that respect the laws of multiple jurisdictions.⁷

⁴ Barbara Endicott-Popovsky and Deborah Frincke, “Embedding forensic capabilities into networks: addressing inefficiencies in digital forensics investigations” (paper presented at the IEEE Information Assurance Workshop, United States Military Academy, West Point, New York, 2006).

⁵ Keyun Ruan, Ibrahim Baggili, Joe Carthy, and Tahar Kechadi, “Survey on cloud forensics and critical criteria for cloud forensic capability: A preliminary analysis” (paper presented at the 6th ADFSLS Conference on Digital Forensics, Security and Law, Richmond, Virginia, 2011).

⁶ Bernd Grobauer and Thomas Schreck, “Towards Incident Handling in the Cloud: Challenges and Approaches” in *Proceedings of the 2010 ACM Workshop on Cloud Computing Security Workshop* (2010), 77–86.

⁷ Keyun Ruan, Joe Carthy, Tahar Kechadi, and Mark Crosbie, “CLOUD FORENSICS,” in *Advances in Digital Forensics VII*, ed. Gilbert Peterson and Sujeet Shenoi (Springer, 2011), 35-46.

Like all “outsourcing,” cloud contracts reflect a decision that certain organization functions can be better provided with resort to non-organizational resources. This means that, ultimately, the services will be rendered by parties that are not employees of the organization and are therefore not within the “direct control” of the organization. As a result, the entirety of the relationship is captured by the terms of the SLA. The SLA is the source of the authority to resolve all issues and disputes between the cloud service provider and the user. If it is not in the contract, it is not part of the formal relationship.

As a result, the terms of the SLA dictate the rights of the user and the duties of the service provider with respect to the availability of forensic data (such as usage logs and other information about the system and its function) for the customer that could be among the types of data that could be required to be produced in the event of a discovery request. If the SLA does not address what type of process or system data (including metadata) will be provided for the customer, then the cloud provider has no contractual duty to provide such information. This creates at least a couple of legal implications: 1) it can result in more limited access by the organization to forensic data than is needed to respond fully to court compelled discovery; 2) it can lower the quality of evidence available to the company.

Importantly, the organization that is the subject of the data request will continue to be responsible to the court to produce the data required. The obligation is not shifted to the cloud provider. As a result, organizations that do not pay sufficient attention to the SLA may find themselves to be caught between a government requirement that they produce data and a cloud service provider that is unable to adequately perform service in that context. Unless the SLA provides otherwise, the failures to adequately respond to the data production requirement can result in penalties, sanctions, and other undesirable results for organizations involved in litigation.

It is important to note that an SLA is only binding between the parties and does not restrict the information that may be sought under a warrant or subpoena issued by a government authority. An SLA that denies access to forensic data requested by the customer (such as metadata) does not provide protection for the customer in the case of a subpoena or warrant from the courts. If a warrant compels production, the provider’s terms from the binding SLA are not a shield for providing additional data such as metadata, log files, or other information. In these cases, if they exist they must be produced.

The rights and duties of organizations under Service Level Agreements and other contracts have broad economic and strategic legal implications, as well as legal impact, if they impair an organization’s ability to perform its legal obligations in a legal case and require that cloud customer to spend more money and resources on eDiscovery production than they might otherwise have to. Organizations should read and review their cloud agreements in order to understand, at a minimum, the additional risks involved with storing information in the cloud. From an information management perspective, legal aspects should be included in analysis of systems to ensure compliance, but also to understand risks involved with cloud integration.⁸ A more extensive discussion of these tradeoffs will be presented in a discussion on opportunity costs.

⁸ M. Farwick, B. Agreiter, R. Breu, M. Häring, K. Voges, and I. Hanschke, “Towards living landscape models: Automated integration of infrastructure cloud in enterprise architecture management” (paper presented at the IEEE 3rd International Conference on Cloud Computing (CLOUD), 2010).

4. Barriers to Usefulness and Admissibility of Cloud-Based Evidence

The field of ‘cloud forensics’ is emerging to address the study of the technical, organizational, and legal issues associated with digital forensics conducted in cloud computing environments.⁹ The field of “cloud forensics” is related to the processes of Electronic Discovery, or eDiscovery.

The term “eDiscovery” refers to the legal requirements that compel organizations to make available all documents and other information relevant in a U.S. civil legal case. eDiscovery refers specifically to “the process of identifying, preserving, collecting, preparing, reviewing, and producing electronically stored information (“ESI”) in the context of the legal process.”¹⁰ These requirements can affect records management in various ways, since they raise the importance within the organization of ensuring an organization’s records are producible, on demand, in an efficient, searchable manner. Planning for cost effective production of ESI from the cloud to respond to eDiscovery requests requires analysis of how characteristics of the cloud environment create unique legal issues.

In general, the eDiscovery process is characterized by six stages, as described by the Electronic Discovery Reference Model (EDRM).¹¹ This reference model is a common framework for the “development, selection, evaluation, and use of electronic discovery products and services.”¹² The first two stages of eDiscovery are (i) Information Management and (ii) Identification, both of which occur prior to any legal action. These two “planning stages” prepare an organization in the event of an eDiscovery request. The third stage, Preservation (including Collection) occurs following a potential litigation hold. The Preservation stage is characterized by a requirement that the organization ensure potentially relevant ESI is preserved. Stage four, Processing/Review/Analysis, encompasses the processing of ESI and potential conversion into a reviewable format; the evaluation of the ESI for relevance; and the continual analysis of the fact-finding process for the ESI. Stage five, Production, involves the producing the ESI into an agreed-upon format to reduce costs and for use in court. Finally, stage six, the Presentation stage is, as the name suggests, the stage at which potential evidence is presented in the court to go through the relevant evidentiary tests such as admissibility, etc., and to be offered as exhibits for the case. These six stages help to describe the general steps of the eDiscovery process, and for present purposes, can act as a reference model for the understanding of eDiscovery, and how its requirements might be affected by the use of cloud services.

The focus on examples from U.S. law are intended to be illustrative, and not comprehensive. It is recognized that the eDiscovery rules will vary from one jurisdiction to another, and the issues raised are intended merely to prompt consideration of steps that can be taken by all organizations, in any country, prior to the institution of litigation or other situation in which information production is compelled, and in which an organizations data handling weaknesses can become more costly (such as through the application of penalties and sanctions) than being a mere embarrassment.

In this article, emphasis is placed on the period prior to the initiation of formal legal action, when the parties, and their respective internal policies and external contracts form the basis for the parties’ data

⁹ Ruan et al., “CLOUD FORENSICS.”

¹⁰ Sharry B. Harris, ed., *The Sedona Conference Glossary: E-Discovery & Digital Information Management*, 3rd ed. (The Sedona Conference, 2010).

¹¹ “EDRM Framework Guides,” *Electronic Discovery Reference Model*, accessed June 25, 2012, <http://www.edrm.net/resources/guides/edrm-framework-guides>.

¹² “EDRM Frequently Asked Questions,” *Electronic Discovery Reference Model*, accessed June 23, 2012, http://www.edrm.net/joining-edrm/frequently-asked-questions#faq_1.

duties and rights. The focus is on the beginning phases of the eDiscovery process in general. Particular requirements of eDiscovery rules in one jurisdiction or another, are not presented.

Both the Information Management phase and the Identification phase present opportunities to focus on planning that can yield significant benefits for the organization. For example, prior to the preservation phase,¹³ the organization should identify all ESI sources that potentially would be required to comply with a “litigation hold”,¹⁴ and in particular those that are stored exclusively in cloud services.¹⁵ The goal is to identify potential barriers presented by the use of cloud services that suggest potential legal implications for future discovery. By understanding and strategizing how to deal with such barriers prior to legal action, an organization can avoid additional costs while mitigating risks.

When there is a “reasonable anticipation” that litigation may occur, a litigation hold must be issued.¹⁶ This means that all ESI potentially relevant to a case must be preserved in the event of a court case.¹⁷ Care must be taken to ensure the document preserved is the same as that that was originally created.¹⁸ This raises questions regarding authenticity which are exacerbated in cloud services where there may be additional costs and burdens imposed on a cloud services user, particularly where their cloud contract is unclear, or where the litigation hold involves voluminous or complex information. This raises the question of whether the rules of evidence anticipate any relief in the cloud context. The Federal Civil Rules of Procedure provide the following exception for ESI that may be inaccessible:

Specific Limitations on Electronically Stored Information. A party need not provide discovery of electronically stored information from sources that the party identifies as not reasonably accessible because of undue burden or cost. On motion to compel discovery or for a protective order, the party from whom discovery is sought must show that the information is not reasonably accessible because of undue burden or cost. If that showing is made, the court may nonetheless order discovery from such sources if the requesting party shows good cause, considering the limitations of Rule 26(b)(2)(C). The court may specify conditions for the discovery.¹⁹

As cloud services become more familiar and ubiquitous, it may be more difficult to assert that information in the cloud is not “reasonably accessible.”

Further, in addition to organization responsibility, legal counsel for an organization has specific duties with regards to eDiscovery. Once a “litigation hold” is in place, legal counsel has a duty to monitor compliance of the production of relevant documents.²⁰

Whatever the specific arrangements entered into by an organization using cloud services, when maintaining and processing information in the cloud, there are certain issues that should be considered to minimize the potential for negative impact. These include: third-party control, jurisdictional issues, and authenticity of data questions. The following discussion highlights these barriers and the unique issues that must be considered for potential cloud-based evidence.

¹³ See next section, “Preservation for eDiscovery”

¹⁴ Ibid.

¹⁵ “Identification Stage, Electronic Discovery Reference Model,” *Electronic Discovery Reference Model*, accessed June 25, 2012, <http://www.edrm.net/resources/guides/edrm-framework-guides/identification>.

¹⁶ *Zubulake v. UBS Warburg LLC*, 2004 U.S. Dist. LEXIS 13574, (S.D.N.Y. 2004) (*Zubulake V*).

¹⁷ *Zubulake v. UBS Warburg*, 220 F.R.D. 212 (S.D.N.Y. 2003). (*Zubulake IV*).

¹⁸ “The Sedona Conference Commentary on ESI Evidence & Admissibility,” *The Sedona Conference*, March 2008.

¹⁹ FRCP 26(b)(2)(B).

²⁰ *Zubulake V*.

5. Jurisdiction and Cloud Computing

In contrast to networked information systems, which apply to the global cloud, nation states are bound to geography. Since 1648, and the signing of the Peace of Westphalia, countries have enjoyed fixed geographic borders that are intended to be respected by other sovereign nations. Within their respective borders, countries are sovereigns that are able to pass their own laws. This results in a patchwork of legal regimes from a global perspective.

The concept of “jurisdiction” in law is a quite formal consideration of whether one legal authority or another (such as a national sovereign) has auspices over a particular legal matter. The tests are varied, depending on context, but typically come down to the idea of whether there is some “nexus” or connection to a jurisdiction.

This patchiness leaves its imprint on the cloud which is deployed on servers and devices around the globe, and more importantly which handles data relating to parties in multiple jurisdictions and subject to multiple separate legal regimes. This simple distributed architecture creates havoc as different courts and regulatory authorities untangle issues of sovereignty and jurisdiction.

Disputes come in many shapes and sizes, so that it is not possible to describe the specific characteristics of a system that are needed to respond to all types of disputes. The spectrum of disputes can range from undiscovered and unasserted nascent claims through full formal legal actions in court, with each stage lending itself to different forms of dispute resolution. Where the dispute results in formal legal action, the actual case can take two forms—civil or criminal:

1. **Civil cases:** where one person or entity brings a claim against another person or entity for a failure of a legal duty. For civil cases, the Federal Rules of Civil Procedure (Fed R. Civ. P.), or similar state jurisdictional rules are followed. Civil cases now typically rely on some aspects of eDiscovery, and the discovery process in Federal courts is followed using the Fed R. Civ. P.
2. **Criminal cases:** which involve criminal charges the government brings under U.S. Code or other state or federal laws. For that subset of such criminal cases that involve computer systems and related systems, forensic evidence is likely to be sought to be admitted as evidence in order to prove a fact. For instance, cases involving unauthorized access to a system will require sufficient evidence to show the break in, and attribute it to a particular suspect.

This article will touch on authenticity issues and other legal considerations involving cloud-based evidence in civil cases, it will not focus on criminal cases. We focus on the eDiscovery process for the production of business records, which are typically produced for civil cases.

In both civil and criminal cases, authenticity of evidence (including forensic evidence relating to a computer system) is critical, yet the field of cloud forensics is just emerging. There are situations in which an unauthorized access dispute may require information from systems not directly involved in the dispute. This potential reason of production may not be anticipated, which highlights the need to ‘bake-in’ discovery support from the very beginning design stages of the information system.

The presence of potential contacts (“nexus”) in multiple jurisdictions is consistently noted as a primary issue in cases involving cloud evidence.²¹ The tests of contacts varies from one jurisdiction to another, and even from one law to another within a jurisdiction. For civil cases, when the data and entities

²¹ Ruan et al., “Survey on cloud forensics and critical criteria for cloud forensic capability.”

involved in the case are in different geographic areas, the primary jurisdictional requirement is that the “forum” state (the state where an action is brought) should have “enough connection with a problem to satisfy constitutional and statutory requirements.”²²

The location of data is obscured by use of cloud services. This raises the question of whether the jurisdiction in which it is stored is even relevant as a place of “connection” sufficient for jurisdiction to be invoked. Storage of information in the cloud is still relatively new, and the law has not caught up to cloud environments. As a result, the answers of law may differ depending on the situation.

For instance, the tax law under the OECD Model Income Tax Treaty Article 5 states that a “permanent establishment” (an income tax treaty term for “connection”) is present if a party owns a server in a jurisdiction, but not if they merely have a website (and presumably data) that is accessible from that jurisdiction. Data breach notice laws of many states, offer a contrasting example. Many of these laws differ from one state to another with the result that the choice of which law applies matters. These laws each provide that it is the law of the jurisdiction in which the affected data subject resides that is applied, not the law of the location of the data and not the location of the company operations that held the data. Thus, companies from which personal information was stolen are often obliged to send out notices that conform to the many different laws where the data that was compromised relates to people from different states. This is only the tip of the iceberg regarding the potential complexity of jurisdiction and suggest challenges for the application of the law for ESI.

One significant issue occurs when the parties and evidence are located in different jurisdictions. A complex area of law called “Conflict of Laws” has been developed to resolve these issues. This body of law is beyond the scope of this article and will not be addressed here. Suffice it to say that there are myriad considerations in different contexts that inform the analysis. The most important point to be made regarding conflict of laws, like jurisdiction issues, is that the issues associated with the multijurisdictional nature of cloud based data storage may be new, but they are not entirely unique, and similar issues arise frequently in disputes outside the realm of cloud storage.²³ What is new in the cloud context is the broader distribution of data, and the ascendancy of data in value and importance, which may in some contexts elevate it to a more significant variable in the consideration of both jurisdiction and choice of law. Stated simply, in the cloud, it is more difficult to determine the “nexus” of activity, and its location, and then understand the proper jurisdiction or applicable law in a case. Current case law offers little overarching direction as to how jurisdictional matters will be solved in eDiscovery disputes involving cloud-based evidence.

6. Third Party Control for Cloud-Based Evidence

One common feature of cloud services is that they demand reliance on one or more third parties that deliver the service. As businesses become increasingly dependent on cloud services, they will require increasing levels of assurance that third party cloud services will be performed in accordance with business, legal and technical standards sufficient to enable the cloud service recipient to engage in its

²² William M. Richman and William L. Reynolds, *Understanding Conflict of Laws*, 3rd ed. (Danvers:LexisNexis, Matthew Bender, 2002).

²³ Aaron Alva, Ivan Orton, and Barbara Endicott-Popovsky, “Legal Process and Requirements for Cloud Forensic Investigations,” in *Cybercrime and Cloud Forensic: Applications for Investigation Processes* (IGI Global, forthcoming 2012).

business in a normalized fashion, and consistent with the expectations of both its management, its shareholders and its customers. In other words, organizations that use cloud services will depend on reliable performance of their cloud service “subcontractors” in order for the companies to perform satisfactorily for their own customers. Cloud providers will be relied upon to provide information-related services consistent with cloud customer needs at each level of dispute; from the very earliest stages of dispute all the way through the rare, but eventful, court case.

At the earlier stages of a dispute, before a complaint is filed in court, the Federal rules of evidence and related court rules relating to formal discovery will not directly apply; but as noted above, they can inform the “planning phases.” In addition, at those earlier stages, sources of dispute resolution rules will vary. They will typically be informed by the agreements, contracts and policies through which the parties receive the cloud services and other services.²⁴ It is recognized that once a formal legal action is initiated, these formal discovery and evidence rules apply, which significantly affects the manner in which the parties interact in sharing information.

Because cloud service users rely on third party service providers and their contracted services, legal questions can arise regardless of any laws or regulations as to the actual possession of the data. Who is the actual custodian of the data? Issues of dominion and control of digital objects for forensic investigations have arisen prior to today’s broader use of cloud computing.²⁵ Now, the networked, multi-tenant, multi-jurisdictional characteristics of cloud computing raise new questions as to which party has control over the data. If the third party retains control over necessary data (such as metadata), then the contract and SLA determine the parties’ obligations that inform how difficult it will be for this data to be recovered. For eDiscovery, these questions arise in the procedures governing the process whereby the discovery request is for “items in the responding party’s possession, custody, or control.”²⁶ The responding party in this case would be the organization in question, though the actual control of ESI may be in the hands of the third-party cloud provider.

Additionally, the rules provide that parties that destroy information may be subject to spoliation penalties. The rapid provisioning of the cloud is a key characteristic that has taken place frequently without consideration of the long-term storage requirements of records management, including issues of records preservation, spoliation and disposal. The supposed exemptions on possession of ESI and on what is considered normal course of business are strongly relevant due to the nature of the cloud, and could be potentially applied to parties that store information in the cloud. Ultimately, the unique aspect of third-party control as a legal barrier for forensics will be strongly dependent on the legal instruments that control the relationship between the cloud provider and cloud consumer.

The early identification phase gives organizations an opportunity to properly plan for and understand the eDiscovery requirements prior to actual legal action.²⁷ By identifying key individuals involved; conducting a data mapping; identifying relevant ESI sources; and certification that all sources

²⁴ This article does not seek to replicate the many articles that discuss the evolution of the rules of evidence and discovery rules to accommodate digital information processing. Readers are encouraged to review those sources for that perspective. *See* Federal Rules of Civil Procedure for eDiscovery, particularly Fed R. Civ. P. 26.

See also Federal Rules of Evidence for evidence admissibility, particularly Fed. R. Evid 901 on authenticity.

²⁵ Michael Losavio, “The Law of Possession of Digital Objects: Dominion and Control Issues for Digital Forensics Investigations and Prosecutions” (paper presented at the First International Workshop on Systematic Approaches to Digital Forensic Engineering, 2005).

²⁶ Federal Rules of Civil Procedure 34(a).

²⁷ *Ibid.*

are valid, an organization can better understand any third-party implications that may arise.²⁸ This includes understanding the level of support to expect from a third-party provider consistent with the language in the contract. Additionally, knowledge of what critical ESI is stored outside of direct control stands as a cost-savings measure for future disputes.²⁹

From a third-party control standpoint, preservation must be conducted for any relevant ESI that is stored in the cloud as well. The fact that ESI is stored by a third-party should not excuse the information from such requirements. Thus, an important area where cloud computing complicates the preservation requirement is with respect to third-party control. For example, an additional step must be taken to inform the third-party that a litigation hold is in place, and that certain ESI must be preserved. If this issue is anticipated in a legally binding agreement between the customer and the provider, then the contract will control such request. In the cases where the cloud provider and customer have no contractual obligations or processes for dealing with preservation, the customer takes on the additional risks and costs associated with the attempt to comply with a litigation hold. These risks and attempts at anticipating compliance challenges and solutions give rise to countless ‘what if’ scenarios that will differ depending on the architecture of the organization’s systems, any records retention policies that may be in place, the willingness of the third-party provider to work with the customer, and more.

Further, there is difficulty for the customer in weighing potential spoliation fines against the higher burden of ensuring evidence stored in the cloud is preserved. The “not reasonably acceptable”³⁰ basis for not being able to preserve evidence would likely be tested to help determine potential spoliation disputes.

8. Authenticity Issues

Authenticity is a critical gate for evidence admissibility in court—to show evidence is what it purports to be. The unique aspects of cloud-based evidence present challenges in authenticity that must be considered throughout the development of processes for storing information in the cloud. A seminal court case on handling ESI advocates strongly for ‘thoughtful advance preparation’ when admitting evidence to court.³¹ Planning to ensure data and the processes to store data in the cloud that are forensically ready would aid in developing a clear showing of admissibility. As such, an understanding how data will be authenticated is relevant.

For a U.S. trial, the prerequisite for admissibility is to show that “the ESI is what it purports to be” (*Lorraine v. Markel*, 2007). There are ten non-exclusive ways suggested by the Federal Rules of Evidence 901 to show authentication:

1. Testimony of a Witness with Knowledge;
2. Nonexpert Opinion About Handwriting;
3. Comparison by an Expert Witness or the Trier of Fact;
4. Distinctive Characteristics and the Like;
5. Opinion About a Voice;
6. Evidence About a Telephone Conversation;
7. Evidence About Public Records;

²⁸ Ibid.

²⁹ See Section 4.

³⁰ Fed R. Civ. P. 26(b)(2)(B).

³¹ *Lorraine v. Markel American Insurance Company*, 241 F.R.D. 534 (D.Md. May 4, 2007).

8. Evidence About Ancient Documents or Data Compilations;
9. Evidence About a Process or System; and
10. Methods Provided by a Statute or Rule.³²

Most potentially relevant for cloud-based evidence are “Testimony of a Witness with Knowledge” and “Evidence About a Process or System.” For these two methods of authentication, an expert witness will (or most likely will for the latter) testify whether the evidence is what it purports to be. The expert witness must lay a foundation qualifying him or her as qualified. This qualification would include the following assertions: “(a) the expert’s scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case.”³³

Proving authenticity by evidence about a process or system is “designed for situations in which the accuracy of a result is dependent upon a process or system which produces it.”³⁴ This may be the case for cloud-based evidence, particularly where cloud-based evidence is gathered by a third-party, the cloud provider. Questions such as attribution of data to a cloud customer (or with further granularity) may be asked in order to show the process for these determinations was well-documented and properly followed.

9. Chain of Possession Issues

The party that is the subject of the information request, i.e., the cloud provider, will typically be less able to affirm that the data was treated or handled in a certain manner simply as a result of it having been held in a cloud service. In this “chain of possession” issue, the cloud service provider is a link in the chain that is outside of the cloud service recipient’s direct control, and as such represents a potential unknown in any data production context, whether or not there was unauthorized access. Clearly organizations using cloud services need to be able to know and attest to the manner in which data for which they are responsible was collected, held, used, transferred and disposed of.

In the seminal case *Lorraine v. Markel*, Magistrate Judge Grimm aptly noted, “the inability to get evidence admitted because of a failure to authenticate it almost always is a self-inflicted injury which can be avoided by thoughtful advance preparation.”³⁵ Such preparation is one of the main purposes of the identification phase for eDiscovery. For cloud-based evidence, admissibility is likely more challenging due to the many technical and legal barriers mentioned in this article. A strategic plan to ensure evidence preserved will contain sufficient ‘data about data,’ or metadata, to properly authenticate it is recommended.

Authenticity issues that arise for eDiscovery will include whether there is sufficient metadata available to confirm the evidence is what it purports to be. As mentioned, the cloud presents many barriers to authenticity, from control of data (availability and chain of possession); to privacy of other tenants; and the use of reputable processes to obtain cloud-based ESI. The implications for eDiscovery will touch on each of these issues, and are not particularly unique to eDiscovery itself. Ensuring proper authenticity will create a structured system whereby effective records management processes will lead to a more cost-effective production in the event of a legal dispute.

³² Fed. R. Evid 901(b).

³³ Fed. R. Evid. 702.

³⁴ Fed. R. Evid. 901(b)(9) Advisory Committee note.

³⁵ *Lorraine v. Markel*, 2007.

In summary, the cloud presents a series of unique issues that should be considered when storing relevant information in the cloud. We do not seek to be comprehensive in this article; there are additional issues that the cloud raises, such as multi-tenancy and its effect on access to shared logs or metadata. This walkthrough raises many questions regarding the particulars of each element that present potential areas for additional research. We also advocate that organizations take the time to understand and plan for a preservation request. While this list is by no means comprehensive, it presents some initial important areas that warrant an understanding of risks involved. This understanding should be informed by the opportunity costs incurred if these issues are ignored.

10. Opportunity Costs

There are clearly tradeoffs involved with storing information in the cloud, and in relying on the cloud for other information processing functionality. The legal implications described in this article are among the factors that should be taken into account in considering the potential costs for choices made to mitigate these problems, as well as cost of choices deferred. The former (i.e., “choices made”) are direct costs associated with design, development and deployment of information systems in accordance with choices made. The latter (i.e., choices deferred”) are opportunity costs, or the relative value of the alternative choice not taken in comparison to the choice that was actually made.

There has been case law setting forth precedent on how the courts will handle costs of eDiscovery, which are relevant in the opportunity costs analysis. The widely used *Zubulake* test has arisen as a legal process for determining such costs.³⁶ The *Zubulake* test arose as an update to an earlier cost-shifting test.³⁷ The previous test, the *Rowe* test, was a standard for decisions in cost-shifting, and was updated in order to provide a more neutral analysis by the courts.³⁸

The *Zubulake* test sets out the following new factors for determining costs:

1. The extent to which the request is specifically tailored to discover relevant information;
2. The availability of such information from other sources;
3. The total cost of production, compared to the amount in controversy;
4. The total cost of production, compared to the resources available to each party;
5. The relative ability of each party to control costs and its incentive to do so;
6. The importance of the issues at stake in the litigation; and
7. The relative benefits to the parties of obtaining the information³⁹

The judge will determine the breakdown of production costs, and will weigh in on cost-shifting. This test offers guidance for what costs may be, though the scope of this article will not consider such costs.

Another potential cost exposure arises with respect to spoliation, which is defined as “the destruction or significant alteration of evidence, or the failure to preserve property for another’s use as evidence in pending or reasonably foreseeable litigation.”⁴⁰ Penalties for avoidable spoliation can be expensive, and the court can even decide that spoliation of important evidence is grounds for ruling. The

³⁶ *Zubulake v. UBS Warburg*, 217 F.R.D. 309 (S.D.N.Y. 2003). (*Zubulake I*).

³⁷ *Rowe Entertainment, Inc. v. William Morris Agency, Inc.* 205 F.R.D. 421 (2002).

³⁸ *Zubulake I*.

³⁹ *Ibid*.

⁴⁰ *West v. Goodyear Tire & Rubber Co.*, 167 F.3d 776, 779 (2d Cir.1999).

Federal “common law” of spoliation creates incentives for organizations to implement systems that protect against loss or corruption of potential ESI.⁴¹ Additionally, statutory or regulatory requirements may impose more requirements for the archiving of records.

Legal instruments that limit eDiscovery present additional risks (and corresponding costs) to an organization based on the language of the contract, and the characteristics of the cloud environment itself. From a contractual perspective, one potential cost is clear. The customer may assume greater costs of production in response to an eDiscovery request if cloud service providers fail to satisfy requests for information production to the satisfaction of the relevant requesting authority. Such costs may be compounded by penalties and fines for failure to comply. Ideally, a cloud customer should negotiate inclusion of the service provider’s responsibility for payment of any penalties and fees that arise as a result of the provider’s failure to produce information in accordance with the terms of the agreement.

Contracts clearly have direct cost implications that could affect the customer. In addition to the third-party control questions regarding who actually possesses and controls the data,⁴² there are also duty-to-preserve questions that could impose costs due to the inaction of the third-party provider.⁴³ At least one source contends that a duty to preserve may extend to evidence entrusted to others, and in such instances “a party may be held liable for spoliation committed by a third party to whom it entrusted the destroyed evidence.”⁴⁴ Thus, a contract also has the potential to directly harm the customer by what it fails to address, such as in those cases where legal preservation of data by the cloud provider is not established as a contractual obligation.

While a lack of explicit forensic and discovery support in a contract or SLA poses a risk to the customer, this opportunity cost is difficult to quantify. More research is required.⁴⁵ Typically, a party will be required to produce ESI even if the costs of production are prohibitive. As such, it is in a party’s interest to ensure an easy, quick and reliable method for evidentiary information retrieval.⁴⁶ Organizations will be served if they view production of ESI as an “ordinary and foreseeable risk” associated with electronic storage.⁴⁷

Organizations should understand these risks and plan accordingly. If an organization stores records electronically, then records retrieval will be part of its ordinary business activities, and compelled records retrieval can reasonably be viewed as an “ordinary and foreseeable risk.” Cloud storage must be viewed

⁴¹ John Christiansen, “Discovery and admission of electronic information as evidence,” in *E-Health Business and Transactional Law, 2010 Cumulative Supplement*, ed. J. Sullivan (Arlington: BNA Books, 2010), 427-52; G. Paul and B. Nearon, *The Discovery Revolution: e-Discovery Amendments to the Federal Rules of Civil Procedure*, (Chicago: American Bar Association, 2006). See also *Silvestri v. General Motors Corp.*, Federal Reporter Third Series, vol. 271, pp. 583–595, 2001.

⁴² See earlier Section “Third Party Control for Cloud-Based Evidence.”

⁴³ Joseph A. Nicholson, “*Plus Ultra*: Third-Party Preservation in a Cloud Computing Paradigm,” *Hasting Business Law Journal* 8, no. 1 (Winter 2012): 191-220, p. 191.

⁴⁴ Margaret M. Koesel and Tracey L. Turnbull, *Spoliation of Evidence: Sanctions and Remedies for Destruction of Evidence in Civil Litigation*, 2nd ed. (American Bar Association, 2006).

⁴⁵ “Results of the 2012 eDSG Investigation of Cloud Service Providers and eDiscovery,” last modified March 7, 2012, <http://ediscoveryconsulting.blogspot.com/2012/03/results-of-2012-edsg-investigation-of.html>.

⁴⁶ Nicolai Kuntze, Carsten Rudolph, Aaron Alva, Barbara Endicott-Popovsky, John Christiansen, and Thomas Kemmerich, “On the Creation of Reliable Digital Evidence,” in *Advances in Digital Forensics VIII: 8th IFIP WG 11.9 International Conference on Digital Forensics Pretoria, South Africa, January 3-5, 2012, Revised Selected Papers*, ed. Gilbert Peterson and Sujeet Sheno (New York: Springer, 2012), 3-18; See also Fed R. Civ. P. 26(b)(2).

⁴⁷ In re BRAND NAME PRESCRIPTION DRUGS ANTITRUST LITIGATION. Nos. 94 C 897, MDL 997 (1995).

as a relevant factor in assessing risk and costs. The framework and discussion of legal control instruments reviewed earlier can inform organizations of these additional risks.⁴⁸

There are clearly costs involved with ensuring systems are forensically and discovery ready. These costs should be compared to those incurred from a “reactive strategy” where little or no planning takes place. In addition, the prevalence of litigation in a particular industry, the nature of litigation in a particular jurisdiction and other similar variables will affect any analysis of risk. Given organization reliance on predictable data systems, it may be in an organization’s financial interests to accept the additional costs of digital evidence creation and retention systems.

11. Conclusion

Prior planning is necessary to ensure forensic and discovery support is ‘baked in’ to an organization’s arrangements for cloud storage and other “outsourced” computing processes, especially in contexts where litigation is prevalent. Further the “data system hygiene” that results can also help to render information systems more effective in ordinary operations.

A particular challenge for cloud service customers is providers’ reluctance to offer explicit or comprehensive treatment of the relative rights and responsibilities of the customer and provider in the context of litigation. As the market develops further, it is possible that explicitly detailed treatment of forensic support by cloud providers will characterize cloud service contracts, but only if cloud service customers demand such services, or if it is compelled by government authorities.

This paper advocates appropriate planning to ensure one’s systems are forensically-ready, and advocates broader understanding of the burdens and benefits that storage in the cloud places on the organization, as well as the third-party provider. The evolution of cloud contract forms will frame that discussion. It remains to be seen what legal recourse will be made available to customers harmed because cloud providers do not comply with customer records requests.

12. Future Work

The current context for cloud computing does not provide clarity regarding an organization’s discovery risks. Further study comparing the risk/benefit of a “planned” strategy and the “reactive” strategy will shed light on the value of forensic readiness. Lawyers must “thoroughly understand the responding party’s computer system, both with respect to active and stored data.”⁴⁹ In previous articles, we have posited that there is a lack of education and knowledge by lawyers (and judges) regarding digital evidence,⁵⁰ and that the gap will grow as systems become increasingly complex. As a result, we will continue our work with legal educators to develop digital evidence education and curriculum that will explore further how to raise the technical literacy of various components of the legal system in eDiscovery and the related legal issues that arise from cloud computing. In addition, we plan to explore toward more specific knowledge in the legal processes surrounding eDiscovery in order to benefit the field.

⁴⁸ Kuntze et al., “On the Creation of Reliable Digital Evidence.”

⁴⁹ *Zubulake I*, 217 F.R.D. at 324.

⁵⁰ Aaron Alva and Barbara Endicott-Popovsky, “Digital Evidence Education in Schools of Law”(paper presented at the 7th ADFSL Conference on Digital Forensics, Security and Law, Richmond, Virginia, 2012).

Models in Collaborative and Distributed Digital Investigation

In the World of Ubiquitous Computing and Communication Systems

Michael Losavio, Deborah Keeling and Michael Lemon

Abstract

Ubiquitous computing and communication systems produce ubiquitous electronic evidence of use in many disciplines. For law enforcement, the use of digital evidence has expanded beyond electronic child exploitation materials into other traditional areas of criminal justice, including homicide, robbery and narcotics trafficking. This expanded utility is also available to any information community in need of historical data. But the growth in the distribution and volume of this information and its storage media create challenges for collection and the validation of reliability. Several ad hoc and distributed models for investigative process may assist both law enforcement and the curation and archival communities. We examine and discuss the data from these models and the future of distributing digital forensic expertise for broad use.

Authors

Michael Losavio teaches on computer engineering and criminal justice issues at the University of Louisville and has taught in Egypt and Mexico. His J.D. is from Louisiana State University.

Dr. Deborah Keeling received her Ph.D. from Purdue University in sociology and is chair of the Department of Justice Administration of the University. Her research interests are in democratic policing in Europe and Asia and electronic crime.

Michael Lemon is a detective for the Bowling Green (Kentucky) Police Department and a M.S. student in digital forensics at the University of Central Florida.

1. Introduction

Ubiquitous computing permeates the world. The number of cell phones exceeds the U.S. population.¹ With the proliferation of digital technology comes a commensurate growth in transactional and content-related electronic information. This creates unprecedented opportunities for the collection of electronic evidence.² Criminologist James Allen Fox of Northeastern University attributes, in part, the 2011 decline in violent crimes in the United States to improvements in digital investigation and electronic surveillance.³

Issues of digital investigation are not confined to criminal justice. One survey of divorce attorneys of the American Academy of Matrimonial Lawyers found that two-thirds of members used the social networking site FaceBook as a primary source of digital evidence for divorce proceedings; the strong

¹ Cecilia Kang, "Number of cellphones exceeds U.S. population: CTIA trade group," *Washington Post Tech Blog*, October 11, 2011, accessed December 20, 2011, http://www.washingtonpost.com/blogs/post-tech/post/number-of-cell-phones-exceeds-us-population-ctia-trade-group/2011/10/11/gIQARNcEcL_blog.html.

² It also presents an unprecedented means of automated surveillance of citizens, an issue of civil liberties and authoritarian oppression of greater and greater importance.

³ Barrett Devlin, "Crime Down Across Nation," *Wall Street Journal (Online)* [New York, N.Y.], 20 Dec 2011.

majority noted an increase in the use of such evidence over the past several years.⁴ And this extends far beyond criminal investigations into civil forensics, data analytics and, indeed, the ways needed to preserve and validate the memory of the world for future generations. Duranti,⁵ Endicott-Popovsky and others have explored the application of digital forensics to the “born-digital” world as a crucial domain for preserving truth in an electronic world. Kirshchenbaum, Ovenden and Redwine have systematically addressed the relationship of digital forensics and born-digital data for cultural heritage.⁶ Digital forensic systems could solve key issues facing archivists with electronic information, such as data recovery and discovery, authentication and accessioning.

Yet, there are shared challenges across all of these disciplines for the effective use of digital forensics in a world of ubiquitous computing. The responses of the law enforcement community may aid all disciplines in meeting them. Law enforcement digital forensics needs have grown with the ubiquity of electronic evidence associated with all types of criminal investigations. Many investigators now look for digital evidence in any case. Either through training, conversations with other investigators or television, use of digital evidence is becoming more and more common in police work. Officers now seek digital evidence as they would surveillance videos, fingerprints and DNA.

One fetal abduction/murder investigation shows the value of digital forensics to any investigation.⁷ The female suspect presented to the local emergency room with a newborn child. When she was examined the ER staff notice the infant had organs attached that should still be in the mother if the mother were still alive. This led to an investigation of what actually occurred.

The examination of the suspect’s computer and cell phone found a scheme to acquire an infant. Evidence from the computer showed contacts with several pregnant females on social media sites and searches of the Internet for how to do a home Caesarean section delivery of a baby. She claimed to help single mothers during their pregnancy. One person seemed to have more contact with the suspect than others; this person and the suspect talked about meeting on the day of suspect’s “delivery.”

The suspect’s cell phone contained the digital trail from that point. Text messages between the pregnant female and the suspect showed her planning to pick up the female and take her shopping for baby clothes. After the planned time to meet, the texting goes quiet for two hours. The suspect then texts her husband pictures of her new baby which she “delivered” in her vehicle.

Eventually she confessed and showed where she had immobilized the victim, bound her and removed the baby from the victim’s abdomen. The digital evidence identified the victim, confirmed that the suspect was not pregnant, located the victim and showed premeditation. This case demonstrates how digital evidence can be interwoven into the fabric of most types of criminal investigation.

But time, funding or access needed for a thorough digital forensics examination are concerns in all cases, whether law enforcement or civil authorities. A two-tier technical problem continues to create

⁴ Margaret M. DiBianca, *Ethical Risks Arising from Lawyers’ Use of (and Refusal to Use) Social Media*, 12 Del. L. Rev. 179, 183 (2011) (citing Am. Acad. of Matrimonial Lawyers, Big Surge in Social Networking Evidence Says Survey of Nation’s Top Divorce Lawyers (February 10, 2010)).

⁵ Luciana Duranti, “From Digital Diplomats to Digital Records Forensics,” *Archivaria* 68 (Fall, 2009): 39-66; Luciana Duranti and Barbara Endicott-Popovsky, “Digital Records Forensics: A New Science and Academic Program for Forensics Readiness,” *Journal of Digital Forensics, Security and Law* 5, no. 2 (2010): 1-12.

⁶ Matthew G. Kirschenbaum, Richard Ovenden, and Gabriela Redwine (research assistance from Rachel Donahue), *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*, Council on Library and Information Resources, December, 2010, accessed July 23, 2012, <http://www.clir.org/pubs/abstract/reports/pub149>.

⁷ Brett Barrouquere, “Kentucky woman sentenced for death of expectant mom,” *Associated Press*, accessed August 24, 2012, <http://www.kentucky.com/2012/03/02/2091640/kentucky-woman-sentenced-for-death.html>.

backlogs in examinations despite the ever-increasing power of forensic computers. Without a solution to these problems the forensic examinations will face continuing challenges as to resources.

First, the increase in the average size of hard drives in the typical digital case is outpacing the speed and processing power of the forensic computers.⁸ What was once considered an extra-large hard drive a few years ago now comes standard on most computers. But the power of computer processors and their ability to examine more data in less time has failed to grow at a similar rate. Any benefit a new forensic computer gives the examiner in its ability to process more data at a faster rate may be lost due to new larger hard drives that take longer to process.

The second issue has been the proliferation and ubiquity of mobile devices. A few years ago a typical case contained one computer and, perhaps, some type of external media. Now the average adult has more than one type of electronic device; a typical case contains multiple items like a laptop computer, cell phone, tablet computer and related media. Each cell phone may potentially be broken down into three pieces of evidence: the cellular device, a SIM card and a SD card. Each item can contain data and must be examined separately. A cellular device overall could have up to 64 GB of data stored locally and 32 GB stored on the external media. The cell device examination alone may address 100 GB of information. Various models of collaborative and distributed digital investigation have been discussed and implemented for addressing this growth in ubiquitous electronic evidence.

One example is the Regional Computer Forensic Laboratory (RCFL) program of the United States Federal Bureau of Investigation (FBI), involving regional collaborations of the FBI with state, regional and local law enforcement to provide centralized, highly trained digital forensic services. The laboratories are designed to provide centralized access to expertise and services in digital forensics. Yet as the demand for digital forensic services has grown, so has the need to develop and expand such expertise to local agencies.

The Internet Crimes Against Children (ICAC) program⁹ and the Secret Service National Computer Forensic Institute program¹⁰ for local law enforcement are other models for distributing digital investigative expertise to local law enforcement agencies for their home implementation. Two RCFLs have experimented with hybrid models that provide distributed digital forensics expertise and tools to local law enforcement agencies supported with the high-level expertise of the RCFL:

1. The FBI's Kentucky Regional Computer Forensics Laboratory¹¹ and the University of Louisville have implemented and are monitoring a distributed digital forensics mini-lab project linked to the state RCFL and using a triage model to allocate resources to cases;
2. The FBI's Kansas City Heart of America Regional Computer Forensics Laboratory, working with forensics software vendor Susteen,¹² has implemented a "Virtual Cell Phone Kiosk" program¹³ that provides local law enforcement statewide access to a shared pool of cell phone forensic software licenses for device analysis.

⁸ Nicole Lang Beebe and Jan Guynes Clark, "Dealing with Terabyte Datasets in Digital Investigations," in *Research Advances in Digital Forensics*, ed. M. Pollitt and S. Sheno (Springer, Norwell, 2005).

⁹ FVTC Internet Crimes Against Children Training and Technical Assistance Program, accessed August 25, 2012, <http://www.icactraining.org/>.

¹⁰ National Computer Forensic Institute, "Class Schedule," accessed August 25, 2012, <http://www.ncfi.usss.gov>.

¹¹ <http://www.krcfl.org/> (accessed August 25, 2012).

¹² <http://www.mobileforensics.com/> (accessed August 25, 2012).

¹³ See www.harcfl.org/Downloads/Documents/harcfl_virtual_cpik_flyer.pdf (accessed August 25, 2012).

These initiatives have developed in an environment of financial stress on law enforcement. Both criminal justice and civil litigators use various *ad hoc* means within the rules of evidence and the rules of criminal procedure and civil procedure to exploit digital evidence. Some of these methods have led to disturbing results, such as in the Julie Amero prosecution by the State of Connecticut,¹⁴ but may also be found to have sufficient indicia of reliability.¹⁵ Review of all these within the context of an evidence regime can help better measure the key issues of reliability and direct the appropriate use of resources. This is true whether for the evidence regime of the United States, Korea or international tribunals.¹⁶

2. The Evidentiary Continuum

2.1 Evidence across a continuum and foundations for reliability

Rules of evidence serve a gatekeeper function to assure a floor of reliability. After that, it is left to a finder of fact whether judge or jury, to assign a particular weight to the evidence and, considering all the evidence together, make factual conclusions to a legally required level of certainty. Digital evidence must meet those requirements of “weight” and probable certainty though it is important to distinguish the levels of certainty needed for different parts of the justice process.

The reasonable suspicion and probable cause requirements for, respectively, an investigative stop and search/arrest warrant, are at the low end of the probability scales. As Carrier has noted, digital evidence may initially be used at that lower probability to obtain a search warrant that itself secures the evidence needed to prove guilt to a higher level of certainty.¹⁷ For criminal prosecutions in the United States, that higher level is defined as “beyond a reasonable doubt;” for civil actions it is defined as “more likely than not.”

The requirements to establish that basic floor of reliability vary with certain types of evidence. Three categories of particular interest to digital investigations are lay evidence, technical evidence and scientific evidence. These are important distinctions, as each requires a higher level of validation and digital evidence be found in each of these categories. Use requirements for this type of evidence may vary with jurisdiction.¹⁸

This distinction as digital evidence was first enunciated by one federal court of appeals in financial prosecution *United States v. Ganier*.¹⁹ The Court observed that “...the categorization of computer-related testimony is a relatively new question,” and found that a digital forensic examiner’s analysis and conclusions from Windows Registry data required special knowledge of computers and forensic software “well beyond that of the average layperson.” This testimony was of “scientific, technical or other specialized knowledge,” not lay witness evidence, and required compliance with the relevant rules of

¹⁴ Robert X. Cringely, “The Julie Amero Case: A Dangerous Farce,” *Infoworld*, December 2, 2008, accessed August 23, 2012, http://www.pcworld.com/businesscenter/article/154768/the_julie_amero_case_a_dangerous_farce.html.

¹⁵ *United States v. Ganier*, 468 F.3d 920 (6th Cir. 2006).

¹⁶ M. Browne, C. Williamson, and L. Barkacs, “The Perspectival Nature of Expert Testimony in the United States, England, Korea, and France,” *Connecticut Journal of International Law* 55 (Fall, 2002): 18; U.N. Doc. IT/32/Rev.7, Part 6, Section 3, Rules of Evidence, Rule 89, (1996), International Criminal Tribunal for the former Yugoslavia, Rules of Procedure and Evidence, entered into force 14 March 1994, amendments adopted 8 January 1996.

¹⁷ Brian Carrier, *File System Forensic Analysis* (Addison Wesley, 2005).

¹⁸ E. P. Imwinkler and P. Giannelli, *Scientific Evidence*, 4th ed. (San Francisco, CA: Mathew-Bender, 2007), p. 2, fn 1.

¹⁹ *United States v. Ganier*, 468 F.3d 920 (6th Cir. 2006).

evidence and procedure to assure those conclusions were reliable. Digital forensic examinations that require the application of special knowledge to draw factual conclusions fall within this domain. When challenged, the witness must be able to establish the reliability per FRE 702.

What sometimes is conflated is that expertise with particular forensic tools may be used to locate evidence, but the fact of the evidence itself is not a conclusion based on that special knowledge. That testimony, such as to the presence of digital contraband, is lay fact witness testimony. The expert evidence may lie elsewhere, such as in an analysis of Registry data leading to the conclusion that the digital contraband or digital evidence was created on a particular time or date via a particular mechanism. It is important to separate issues relating to technical systems that find evidence from technical systems that produce conclusions that themselves are evidence and must meet the standards of FRE 702.

Further, technical expert evidence and scientific expert evidence may be separate domains that also become conflated. These distinctions may be seen in comparing digital forensics as generally practiced and computational forensics, where computing systems are used to derive factual conclusions in a variety of areas. Digital forensics systems generally serve to find evidence, acting as pointers to that which is presented to the forum. Computational forensic systems, such as information retrieval or data mining tools, may similarly point to the actual evidence. But they may also derive conclusions, such as to chemical or genetic composition, where the conclusions are the evidence.

Lastly, rules of procedure, or how a forensic inquiry proceeds in a court of law, differ in civil actions from criminal prosecutions and differ from one jurisdiction to the next. Criminal prosecutions place a greater burden on the prosecution in that it must establish the entirety of its case on its own. Civil actions permit the parties access to each other so as to ease the burden of proving the facts at issue; they also permit greater sanctions where one party fails to cooperate in the fact-finding process. For digital investigations, parties in civil proceedings may, in effect, be able to require the opposing party to help establish facts at issue relating to digital evidence or face sanctions for not doing so. But in criminal matters, the prosecution must usually carry that burden in its entirety.

3. Review of Distributed Models

3.1 U.S. F.B.I. Regional Computer Forensic Laboratory (RCFL)

The Regional Computer Forensic Laboratory (RCFL) program of the United States Federal Bureau of Investigation (FBI) promotes regional collaborations of the FBI with state, regional and local law enforcement to provide centralized, highly-trained digital forensic services.

The federal government provides funding for the regional facilities and training for staff; in turn, state and local agencies staff the laboratories. The facilities meet classified standards and are equipped with digital forensic systems vetted by the FBI. Staff are trained to FBI standards for the examinations of devices. In turn, the RCFLs accept cases from their service area agencies for analysis in both federal and state prosecutions. The RCFLs also offer examination facilities, particularly cell phone kiosks, for local officers to use as they bring devices to the labs and examine the devices themselves.

3.2 Internet Crimes Against Children (ICAC) program

The Internet Crimes Against Children (ICAC) program distributes digital investigative expertise through specialized training and support for state task force development.²⁰ It is a national collaboration of 61 task forces representing over 2,000 federal, state and local agencies. The focus is on cyber enticement and child pornography cases and is a response to the growth in children and teenagers using the Internet. The Office of Juvenile Justice and Delinquency Prevention asserts that, since 1998, ICAC has supported investigative training for more than 338,000 officers in the U.S. and 17 countries and its Task Forces have been involved in the arrest of more than 30,000 individuals.

3.3 U.S. Secret Service

The Secret Service training and equipment program for local law enforcement is coordinated through its National Computer Forensics Institute.²¹ The NCFI provides classrooms, mock court, computer forensic laboratory and other facilities for training state and local law enforcement from across the United States. Its programs range from several days to several weeks in length and topics from basic computer evidence collection to network intrusion response. Travel, lodging, per diem and training expenses are all paid by the NCFI, making its programs available even to police departments in financially-challenged jurisdictions. After the training law enforcement officers are given the equipment, software, toolkits and manuals for conducting computer and electronic forensic examinations. The NCFI program also includes prosecutors and judges in its training program, creating a broad, coherent expertise on computer forensic and digital evidence issues within the system of criminal justice in the U.S.

These programs have served key roles in distributing expertise and, with the FBI program and ICAC Task Forces, offering central support for ongoing activity.

4. New Collaborative and Distributed Models

4.1 Kentucky Regional Computer Forensics Laboratory - University of Louisville Digital Mini-Lab Project

The University of Louisville has implemented and monitors a distributed digital forensics mini-lab project linked to the Kentucky RCFL and using a triage model to allocate resources to cases. This established and supports digital forensics “triage” stations around Kentucky as part of a larger effort by the Kentucky Regional Computer Forensics Laboratory (KRCFL) and the University of Louisville to make computer forensics more easily available to law enforcement in Kentucky, with no law enforcement agency in the state more than 90 minutes from a facility. While the triage sites or “mini-labs” collaborate with the Kentucky Regional Computer Forensics Lab, they are not organizationally linked with the RCFL program of the FBI and are funded through a grant from the U.S. Department of Justice, COPS Technology Program.

²⁰ Internet Crimes Against Children Program, Office of Juvenile Justice and Delinquency Programs, U.S. Department of Justice, accessed December 20, 2011, <http://www.ojjdp.gov/programs/progsummary.asp?pi=3>; FVTC Internet Crimes Against Children Training and Technical Assistance Program, accessed December 20, 2011, <http://www.icactraining.org/>.

²¹ United States Secret Service, National Computer Forensics Institute, accessed December 20, 2011, <http://www.ncfi.usss.gov/>.

Participating agencies include the Bowling Green, Kentucky Police Department, the Owensboro, Kentucky Police Department, the Paducah, Kentucky Police Department, all in western Kentucky, and the Ashland, Kentucky Police Department and the Kentucky State Police, Frankfort, Kentucky, through its posts in Hazard, Kentucky (Hazard Post 13), Pikeville, Kentucky (Pikeville Post 9), London, Kentucky (London Post 11) and Morehead, Kentucky (Morehead Post 8) all in eastern Kentucky. These agencies agreed to collaborate to expand the capacity for handling electronic evidence in order to enhance public safety in the Commonwealth of Kentucky.

Each participating agency or agency division provides:

1. A secure space with utilities for the location, operation and use of the hardware/software systems;
2. A detailed employee, sworn or unsworn, to be trained and use the systems;
3. Information on other employees to be trained in these forensic techniques; and
4. Availability to accept some cases from other jurisdictions in their use of digital evidence.

The University of Louisville and KRCFL provide:

1. A digital forensics examination system of computer hardware;
2. A digital forensics suite of examination software (AccessData's FTK and MPE suites);
3. A connection to the DCAP network linking each facility to the KRCFL in Louisville
4. Training on these systems and software; and
5. Ongoing support from the KRCFL staff.

This project is built around a triage model for computer/digital device examinations, with standard examinations being handled locally and highly-technical problems being handled by the KRCFL.

4.2 Operational Protocols for the Minilab Model

The KRCFL minilab model requires adherence to a set of rules—protocols of the participating agencies to build a collaborative network to cover the entire state. Those are:

Protocol 1 – relating to a secure space with utilities for the location, operation and use of the hardware/software systems

The equipment, software, documentation and any evidence is maintained in a secure, locked space with utilities for the location, operation and use of the hardware/software systems; these is provided by the Examining Agency at its own expense. Individuals not trained in forensic examination shall not use the machines although they may observe operations and use of the machines under the supervision of a trained examiner. The equipment and software may not be used for any purposes other than the forensic examination of evidence.

Protocol 2 – relating to detailed employees, sworn or unsworn, to be trained and use the systems

The examining agency may use any employee of the department, sworn or unsworn, to conduct computer forensic examinations once that employee has undergone the basic training for such examinations, including FTK training. It will not allow untrained employees to use the equipment for any purpose and will not allow trained employees to use the equipment for any purpose unrelated to official law enforcement purposes.

Protocol 3 – relating to availability and acceptance of and reports on cases, including those from other jurisdictions in support of law enforcement’s use of digital evidence

Examinations of evidence from case agents of the examining agency shall be processed according the investigative and reporting procedures of the examining agency, subject to the data collection requirements of UOL that the examining agency quarterly provide, among other information, copies of log sheets, the Form A, the Service Request Form, and Form B, the Preliminary Examination results form, for its cases where examinations for electronic evidence were conducted by the examining agency using equipment, training or other resources provided under this project. The examining agency, at its discretion, will accept digital evidence for examination from agencies in their and other jurisdictions subject to the availability of resources and compliance of the submitting agency with these protocols and the requirements of the examining agency.

Examinations of evidence from case agents of other submitting agencies shall be conducted as follows:

4.2.1 Contact prior to submission

Prior to presenting evidence for examination, a submitting agency will contact the examining agency and schedule a date and time for an examination of the evidence to be done in the presence of the submitting agency’s case agent.

4.2.2 Submission

All items submitted for evidentiary examination must be accompanied by a completed Service Request Form that includes information and documentation of:

- a. Documentation of legal authority to search
- b. Submitting agency information
- c. Services requested
- d. Incident/Suspected Criminal Activity
- e. Items to be searched

4.2.3 Acceptance procedures

Upon completion of the submission requirements, acceptance of the case for examination by the examining agency and scheduling of review appointment, the submitting agency case agent will transport the evidence items to the receiving agency for review. The case agent will keep the items in his or her custody during the examination, unless the examining agency, at its discretion, determines otherwise to conduct the examination. Each accepted case is logged on a sheet detailing. Submitting agency, submitting case agent, item, time, date, receiving examiner, summary of exam results and time/date item/case agent.

4.2.4 Examination

The examiner will conduct the examination in the presence of the case agent unless otherwise noted. Upon completion of the examination the examiner will complete the Preliminary Examination form, Form

B, and provide it to the case agent with a copy for the examiner's file and copy to be reported to UOL. That report shall detail:

- a. Person/agency submitting evidence
- b. Process for preserving evidence
- c. Process for examining evidence
- d. Results of examination
- e. Verification by and of examiner
- f. Disposition of the evidence.

Where another agency has submitted evidence for examination, the receiving department may require the presence of the custodian of the submitting agency to be present during the examination and to receive back the evidence upon completion of the examination. The submitting agency is responsible for submitting suspected contraband to the National Center for Missing & Exploited Children's Child Victim Identification Program (CVIP)

4.2.4 Special Circumstances

Where it appears the examination may require deposit of the evidence with the examining agency, that agency decides at its discretion whether or not to accept custody of that evidence and shall comply with standard procedures for the preservation of the chain of custody and evidentiary integrity of that evidence. The examining agency may choose to make a mirror- image, bit copy of the submitted evidence with custody to remain with the submitting agency and conduct its examination on the bit copy. The submitting agency should acknowledge the examining agency has no obligation to retain the bit copy and it remains the submitting agency's responsibility to preserve its evidence. Where the examining agency finds additional issues relating to the evidence it may refer the submitting agency to the KRCFL or KSP for further examination.

4.3 Results

The data below shows the use of the mini-lab tools by the different agencies with additional information such as numbers of computers and numbers of cell phones examined, time periods and types of devices.

4.3.1 Agency A

Agency A did not have digital forensics capabilities prior to the project but provided the most complete data on the implementation of the project; their data indicated, inter alia, the expanding usefulness of cell phones as evidence sources in more and more traditional areas of law enforcement. The data is set out in Table 1.

Agency A reported individual examination request forms that detailed requests relating to investigations of murder, robbery, rape, unlawful imprisonment, assault, narcotics trafficking, child pornography and child sexual exploitation.

The majority of the September, 2011 cases (three of five) were related to child pornography/exploitation. By contrast, the majority of the June, 2012 cases (four of five) dealt with other kinds of cases, including homicide, rape and robbery. This indicates the shift towards use of digital forensics as an investigative tool across the criminal justice spectrum.

Table 1.

September 2011 - July 2012	# Computers	# Cell phones (SIMs included)	Other²²
September		6	
October	5	18	1 ²³
November		5	
December		7	
January		2	1
February	3	13	7
March	2	10	3
April		9	4
May	1	4	3
June	2 ²⁴	9	2
July	1	5	2
totals	14	88	22

The number of devices examined in relation to each investigation ranged from one to 12, ranging from computers and cell phones to GPS systems. Half of investigative examinations in the first quarter, 2012 were of multiple devices. Cell phone examinations included SIM card examination. Multiple agencies used the services of Agency A, including the federal Bureau of Alcohol, Tobacco and Firearms (ATF) and the Kentucky State Police.

4.3.2 Agency B

Agency B had a pre-existing digital forensics program which incorporated the new tools into their operations. The reported data are set out in Table 2.

Table 2.

Types of Devices Examined	Numbers examined in time period		
	1/1- 6/13/2012	2011	2010
Computer/HDD	35	56	70
Cell phones/SIM cards	86	67	62
Other (usb drive, CD/DVD,SD cards, diskettes, GPS)	28	210	189
Internal v. External Agency Examinations			
Internal Examinations	77		
External examinations for other agencies	49		

This data further supports the trend away from general computers as sources of digital evidence to mobile devices such as cell phones. Agency B also performed a significant percentage of its examinations for

²² MicroSD cards are counted as separate devices although they are associated primarily found associated with cell phones in these examinations.

²³ TomTom GPS 1EX00.

²⁴ Includes a tablet computer.

other law enforcement agencies in the area. Although case type data was not available for this agency, it did handle the digital forensic examination for the fetal abduction/homicide case discussed above.

4.3.3 Agency C

Agency C had a pre-existing digital forensics program that absorbed the new equipment and training within that framework. Agency C also has an ongoing schools program for discussing online safety issues with school children. With their systems they examined approximately 48 devices in 15 to 20 different cases. The ratio of computers to mobile devices examined was about 1:1, but the agency noted the trend was strongly towards more mobile devices than computers. While the majority of cases continued to be child pornography, examinations were done relating to drug trafficking, counterfeiting, forgery and a car bombing.

4.3.4 Agency D

Agency D had no prior digital forensics operations but used the training and equipment to begin these examinations as well as make the examination systems available to other local law enforcement agencies in their region. Despite schedule conflicts that developed,²⁵ it was able to submit usage information for the collaborating sheriff's office in the county (Table 3).²⁶

Table 3.

Case Number	Agency	Case Type	Date	Media Examined
2012-9107	Daviess County Sheriff's Office	Child Pornography	2/2012	Motorola Atrix
2012-9107	Daviess County Sheriff's Office	Child Pornography	2/2012	Samsung Galaxy
2012-9107	Daviess County Sheriff's Office	Child Pornography	2/2012	HTC
2012-9107	Daviess County Sheriff's Office	Child Pornography	2/2012	HTC
09-64661	Owensboro Police Department	Child Pornography (Practice)	5/2012	Toshiba Satellite A205 Western Digital: WD5000BEVT 500 GB
12-050743	Owensboro Police Department	Child Pornography	6/2012	Custom built PC Western Digital: WD3200AAKS-00L9A0 320GB

4.3.5 Agencies E, F, G & H

Four agency offices were not able to implement the program. In one the officer trained on the systems retired. In another, the trained officer was transferred to a central office digital forensics unit. In all four

²⁵ The police officer assigned to the project was activated for military duty.

²⁶ Report of Cheryl Purdy, instructor in IT and Computer Forensics at Owensboro Community and Technical College, special deputy with the Daviess County Sheriff's Office and on loan to Owensboro Police Department one day per week for the purpose of digital forensics examinations.

offices it appeared the trained officers were needed to continue regular policing duties and did not have full opportunities to provide digital forensic services. Plans are underway to reallocate these resources as needed for supporting digital forensics programs.

4.4 The Data Trends in Digital Investigations

The most significant trends seen in the data are:

1. The significant growth in the numbers of cell phone examinations while;
2. Numbers of general purpose computer examinations remain stagnant or are declining;
3. The number of devices examined in relation to a single investigation may vary but seem to be increasing;
4. Digital forensics is being used for more and more different types of criminal investigations as devices are found at crime scenes and collected by investigating officers; the reported data of Agency A used digital forensics in murder, robbery, rape, assault and narcotics trafficking investigations as well as those for child pornography and child sexual exploitation.

4.5 Administrative Issues

In general, those agencies that have been the most productive under this program are those in which there existed:

1. Adequate manpower resources to reallocate assignments to move field officers into the digital forensics lab part- or full-time;
2. Officers interested in developing the appropriate skills and qualifications to conduct exams, despite the challenging material, and
3. A willingness to promote services with proximate agencies and take on evidentiary examinations from outside their agency.

4.6 Heart of America/Kansas City Regional Computer Forensics Laboratory Virtual Cell Phone

Kiosk Project

The HARCFL Virtual Cell Phone Kiosk Project is a collaboration with Susteen, Inc., the vendor of the Secure View cell phone forensics examination system. It is a response to the explosive growth in cell and smartphone use in relation to crime that pushes analysis tools to more local police departments. The system consists of the Secure View software package, a phone cable kit, an online license validation process (a virtual “dongle”) and an online training program for sworn law enforcement officers; the local department must have its own computer for the use of the system. It is designed for straightforward use by all law enforcement with the evidence retained on the local machine. Secure View 3, the forensics suite currently used, provides for data acquisition, analysis and reporting relating to cell phone examinations. It offers keyword searching and transactional timeline, frequency and linkage visualization for phone, text and web activity on the device.²⁷

²⁷ Secure View 3 Case Management-Analytics, accessed July 9, 2012, <http://www.mobileforensics.com/svProbe>.

4.6.1 Operational Protocols for the Virtual Cell Phone Kiosk Project

The Virtual Cell Phone Kiosk is available only to sworn law enforcement officers. It may be used for all examinations of legally seized or possessed cell phones relating to crime except for child pornography; child pornography examinations must be done either by the HARCFL or at its laboratory cell phone kiosk. An eligible agency applies online with Susteen,²⁸ giving information about the individual and agency applying. This is forwarded to HARCFL for review and approval. Once approved, Susteen creates an account for that user and forwards them the log-in information. The user logs-on to download and install the Secure View forensic software on the local machine to be used for examinations. The agency has the option of purchasing a regular or extended set of cell phone cables to connect the device under examination to the examining machine; it may also choose to purchase those cables from other vendors. Online training in the use of the system is conducted twice weekly for about two hours each session. HARCFL supports several licenses for Secure View. To conduct an examination, the user logs-on to the project site. If a license is available for use, the system validates the use of the local machine and software for that examination session. This, in effect, allows easy sharing of the forensic tools without each local law enforcement agency having to purchase the license.

For federal fiscal year 2011, the first year of operation, 69 agencies had signed up and acquired the Secure View tool, using it for 914 data acquisitions or attempts.²⁹ This may be compared to the 1,610 logged events at the preexisting cell phone forensic kiosk physically located at the RCFL. The virtual tool removed the need to travel to the RCFL with the commensurate loss of officer time. For the first six months of FY 2012, under a revised and narrowed set of definitions, the system logged 442 complete cell phone acquisitions; this did not count use of the software for analysis nor attempted acquisitions that were not successful.³⁰

4.6.2 Data and Data Trends in Digital Investigations with the Virtual Cell Phone Kiosk

Log data for the most recent period of June 5, 2012 through July 4, 2012 show significant use of this system by local departments. Although data entry was inconsistent, the following estimates on system use for that one-month period are:

- Total successful exam logins – 210
- # agencies using system – 40
- # different mobile devices – 74
- # different manufacturers – 12

The most frequently found systems were Apple, HTC, Motorola and Samsung devices, usually cell phones although iPads and Galaxy tablets were examined. Chinese ZTE cell phones were also examined. The law enforcement agencies using the tool included municipal, county, federal and state law enforcement, including the Office of the State Fire Marshall, Lincoln University Police Department and the Kansas Department of Wildlife and Parks. The total count of devices was 249 for the period, although this number must be further verified. Some data entry could repeated listings for identical types of devices

²⁸ Registration site for HARCFL – Susteen Online Cell Phone Kiosk Program, accessed July 9, 2012, http://secureview.us/SV3_HARCFL/.

²⁹ FY 2011 Annual Report, Regional Computer Forensics Laboratory Program, Federal Bureau of Investigation, U.S. Department of Justice, pp 24 - 25.

³⁰ HARCFL presentation on system operations.

such there may have been repeated entry for the same device. But other acquisitions did not list the devices, indicating a possible undercount.

Identifiable crimes were entered for 23 % (48 out of 310) of the examinations. The types of crimes associated with these examinations were:

- Homicide
- Aggravated rape
- Aggravated assault
- Narcotics offenses

The most frequent cell phone examinations were related to drug offenses, followed by sex offenses, homicide and assault. The system has been adopted by the Missouri Internet Crimes Against Children Task Force to provide these services to Missouri law enforcement officers.³¹

This data confirms the need for capacity for cell phone examinations and the expansion of the use of digital forensics to all types of criminal investigations, including those by officers of the State Fire Marshal and Wildlife and Parks. It also indicates the potential growth in the need for examination capabilities for tablet computers. The quick adoption of this tool by local law enforcement indicates it has addressed an unmet need of local law enforcement.

5. The Future of Distributed Models and Conclusion

The data collected from the tests of the two distributed models show the growth in the use of digital forensics in more and different types of criminal investigations. Given the costs of forensic tools and the economic pressures on local governments, these distributed models offer immediate aid for them. The scope and depth of investigation may not be as great as with fully equipped and staffed digital forensics laboratories, but a basic level of essential service is provided. This suggests the benefits of a continuous services model for addressing digital investigations.

Continuous layers of expertise and services for analysing digital evidence can effectively and efficiently expand law enforcement capabilities. It begins with basic police investigative skills in digital evidence and links that seamlessly through to digital forensic expertise and computer science analysis.

In this global model, there is more than a simple distribution of expertise. Rather, with the distribution of continuous levels of skills in identifying and collecting digital evidence there is also a collaborative association of all the agencies within this continuum. As a participating distributed agent finds issues beyond her training, she may immediately consult peers or expert assistance at the RCFL or FBI/Quantico/DHS level.

The system of justice is a highly collaborative and distributed system that relies on both local and national efforts. Digital crimes involving electronic evidence are no different. To successfully provide for public safety in this era of ubiquitous computing and communication we must define and implement systems for the all forms of misconduct. This requires the normalization of digital investigation for all levels of law enforcement and the preparation of the system of justice for properly processing that information. It is the only way to assure public safety while protecting the civil liberties inherent in a culture built on the free and open exchange of ideas.

³¹ Missouri Internet Crimes Against Children Task Force Online Mobile Phone Kiosk Program, accessed July 9, 2012, http://www.secureview.us/MOICAC/SV3_MOICAC_LandingPage.html.

Cloud Computing Implications to Digital Forensics

A New Methodology Proposal

Fabio Marturana and Simone Tacconi

Abstract

This paper deals with a novel approach to digital investigations, aimed at optimizing law enforcement's tasks, concerning digital evidence acquisition, examination, analysis and reporting, and reducing investigation complexity and operational costs. In the face of Internet's pervasiveness and massive market penetration of high-performing and low-cost handset devices, resulting in a worldwide diffusion of cybercrimes, digital forensics seems indeed to be facing severe challenges which, if not taken seriously, may rapidly jeopardize the achievements of many years of forensic research. Motivated by this, the author proposes a model to perform complex data processing which exploits cloud computing capabilities, on the one hand, and modern mobile handsets capabilities, on the other hand, which can be exploited on the crime scene to perform live and post mortem artefact analysis. The paper provides the reader with the design guidelines, architecture components, interfaces, functional and non-functional requirements of a cloud-based forensic platform implementing the so-called "forensics-as-a-service" delivery strategy. The proposed framework consists of a cloud-based server-side and a client side, running on an Android smartphone or tablet. The client-side, in particular, consists of a collection of open source forensic tools and an Android software application, specifically developed for the purpose.

Authors

Fabio Marturana is a Ph.D. student at Computer Science Engineering Faculty of University of Rome "Tor Vergata". He holds a M.S. in Electronic Engineering from Polytechnics of Turin. His research interests include computer and mobile forensics, information security and cloud computing. In these fields he is co-author of the book "Cybercrime and Cloud Forensics: Applications for Investigation Processes", IGI Global Publisher and of several scientific papers in international conferences.

Simone Tacconi holds a Ph.D. in Artificial Intelligent Systems from the Polytechnic University of Marche and a M.S. in Electronic Engineering from University of Ancona. He is currently Technical Director of the Computer Forensics Unit at the Italian Postal and Communications Police. His main scientific interests include computer security, cryptography, cloud computing and digital forensics. In these fields he is co-author of the book "Cybercrime and Cloud Forensics: Applications for Investigation Processes," IGI Global Publisher and of several refereed scientific papers in international journals and conferences.

1. Introduction

During a digital investigation, forensic analysts are used to performing queries, indexing data, calculating hashes, extracting features and correlating partial data to narrow the search and solve the case. As long as investigation complexity, on the one hand, and amount of data to analyse, on the other, allowed it, forensic data processing occurred in sequence on stand-alone forensic workstations. Unfortunately, Internet's pervasiveness and market availability for cheap and sophisticated mobile devices with large storage capacity, have changed the landscape, resulting in an increase of digital investigation complexity and contributing to the global diffusion of cybercrimes as well. Such crimes, on the one hand, are evolving at an astounding pace, following the same dynamic as the inevitable penetration of computer technology and communication into everyday life. Whilst society is inventing and evolving, at the same

time, criminals are deploying a remarkable adaptability in order to derive the greatest benefit from it. Digital forensics, as a consequence, seems to be facing new challenges which, if not taken seriously, may rapidly render the actual forensics techniques obsolete and even not practicable. Current trends in computing and communications technologies have indeed shown a growing amounts of disk storage and bandwidth available to ordinary computer users, resulting in significantly larger forensic data, with regards to the ability to process them in a timely manner. Performing common forensic tasks, such as keywords indexing and image thumbnail generation against a captured image, indeed, may take much of the available time before an investigation can even begin. As a consequence, digital forensics practitioners, attempting investigations using a single workstation as a platform, will be very soon completely overwhelmed by backlogs. A possible solution to overcome the issues outlined above is then to develop a new forensic paradigm exploiting capabilities offered by cloud computing. New discussions are indeed emerging, among forensic practitioners, on whether the cloud ecosystem could be adopted or even extended to delivery law enforcement's forensic tasks "as-a-service". Cloud computing and Internet's pervasiveness have indeed radically changed the way information technology services are created, delivered, accessed and managed. Cloud computing has innovated information technology, enabling tasks formerly carried out by well-rounded computers and servers to be performed on a mobile device such as a smartphone. This new service delivery paradigm has the potential to become one of the most transformative developments in the history of computing, so why not use this technology in your favor then? . Being available on the cloud and being independent of the device used, indeed, a pool of available resources such as applications, processes and services can be rapidly deployed, scaled and provisioned, on demand. For this reason using cloud computing to deliver on demand forensic services could be an effective new way to support digital investigations, allowing cloud organizations, service providers and customers, to establish forensic capabilities and reducing cloud security risks.

In this regards, the paper's aims at describing the design guidelines of a secure forensics-as-a-service delivery platform exploiting cloud computing and mobile devices capabilities to support live and post mortem investigations on the crime scene. The platform is designed to provide investigators with a wealth of computing capabilities to conduct a live analysis on the crime scene with the stand-alone client device. The same device may be used to connect to the cloud platform as well, in order to upload forensic data to the server-side for remote processing.

The remainder of the paper is organized as follows: Section 1 describes the proposed platform, the server-side architecture, the functional and non-functional requirements and the client-server interfaces. Section 2 deals with the architecture and the functional requirements of the client-side, running on a mobile device and then conclusion and possible future research directions are discussed in the last two sections of the paper.

2. The Server-Side

To provide a viable solution to the above outlined issues and following on previous work by Marturana et al., the author proposes a model for distributed forensic data processing, called "forensics-as-a-service delivery platform," aimed at exploiting cloud computing capabilities. The proposed solution assumes that forensic images captured by the client-side software, may be divided into smaller fragments and uploaded in parallel to the server-side platform to manage massive uploads more efficiently. The server-side will be responsible for acknowledging the fragments received and forwarding them to selected servers in the

cloud. Exploiting distributed file system's capability to replicate files, image fragments are then exchanged among the selected servers, and more copies of the captured image are then rebuilt to assure redundancy.

To allow forensic tasks, such as keyword searching, to be split and performed in parallel on different servers, the author have indeed made the assumption that a forensic image can be considered as the natural unit of storage and processing. It is indeed possible to run remote processes on a set of distributed commodity servers capable of caching in RAM and processing image files, allowing complex forensic tasks to be quickly performed in parallel. In this scenario few dependencies place constraints on the distribution of files or require complex distributed synchronization.

2.1 Service architecture, functional model and interfaces

The main architectural components of the server-side, summarized in Figure 1, are: a central process, called Orchestrator (O), a number of remote Forensic Processes (FP), a communication interface (i.e., CS-to-SS) between the client-side and the server-side of the platform, for service request and result delivery, and an internal interface (i.e., O-to-FP) for data exchange among the cloud servers.

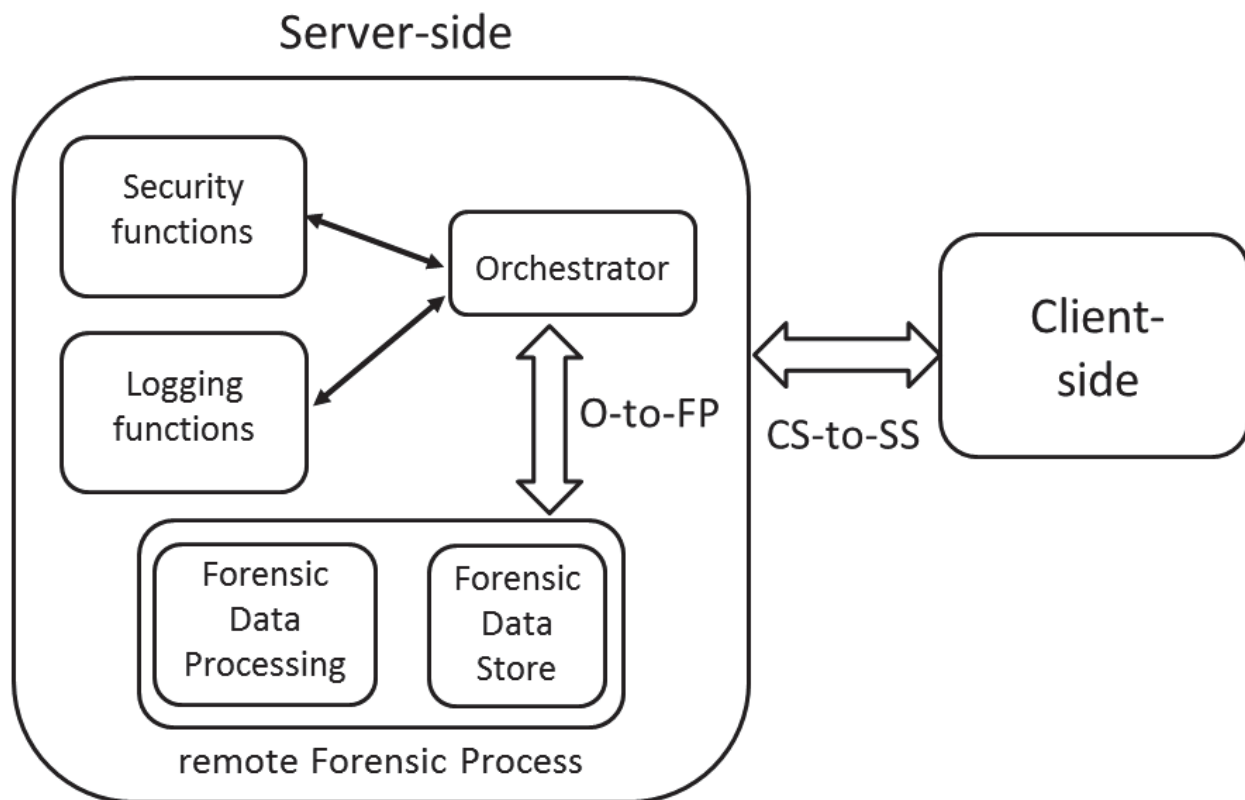


Figure 1. Service Function Model.

On the CS-to-SS interface, the Orchestrator will be in charge of:

- Accepting incoming service requests;
- Initiating outgoing communications;
- Aggregating and delivering results;

while on the O-to-FP interface, it will be in charge of:

- Distributing captured image fragments, once uploaded, among available remote Forensic Processes;
- Keeping trace of captured images retained by each remote Forensic Process;
- Distributing processing tasks among available remote Forensic Processes, with regards to retained images;
- Sending flush command to remote Forensic Processes for deleting retained forensic data when needed.

Remote Forensic Processes will be responsible for:

- Receiving, acknowledging and storing captured image fragments;
- Processing captured images;
- Deleting retained images and forensic data upon Orchestrator request;
- Delivering results to the Orchestrator.

2.2 Server-side functional requirements

Forensic data processing requests and output delivery requirements of the proposed forensics-as-a-service delivery platform are described as follows:

- *Upload of a “live” digital media logical acquisition to the server-side.* The author considers here a live forensic scenario in which a logical copy of all the powered-on digital devices (i.e., computers, tablets, mobile phones, smartphones, PDAs etc.), found at the crime scene, is transmitted to the server platform. A secure connection available at the crime scene (e.g., a VPN tunnel upon a Wi-Fi or a 3G data link) shall be used to upload images. Alternative data transfer procedures shall be considered in case the crime scene is not under network coverage, such as the acquisition of a forensic images at the scene and consequent upload, once back at the forensic lab.
- *Upload of a “post mortem” digital media logical or physical acquisition to the server-side.* In this scenario, a post-mortem investigation on digital artefacts is conducted in a forensic lab. Physical and logical content of such digital artefacts shall be acquired and uploaded to the server-side platform. A secure network connection available at the lab shall be used to upload images. Once uploaded, data shall be retained by the platform for the whole duration of the investigation and may be queried at any time.
- *On-demand analysis of the information uploaded to the server-side.* Once uploaded to the platform, it shall be possible to access and process forensic images at any time. The following is a list of possible operations that shall be executed:
 1. *OS information retrieval:* it shall be possible to garner information about the digital artefact’s operating system;
 2. *Deleted file retrieval:* it shall be possible to use carving tools to extract the list of deleted files from unallocated space and slack space;
 3. *File classification by category or extension:* it shall be possible to classify the content of the analysed image on the basis of statistics on file type, metadata, extension, average dimension etc.;

4. *Installed application software list retrieval*: it shall be possible to retrieve the list of installed software;
5. *Web browsing cache, cookies and history repository analysis*: it shall be possible to extract web browsing evidence, concerning navigation cache and history, list of stored cookies, and related to different browsers (e.g., IE, Firefox, Chrome, Opera, Safari etc.);
6. *Saved chat/IM communication retrieval*: it shall be possible to retrieve evidence concerning saved chat and instant messaging sessions, if stored locally;
7. *Saved Skype calls, chat and messages*: it shall be possible to retrieve evidence concerning saved Skype calls, chat and messaging sessions, if stored locally;
8. *Local email database extraction*: it shall be possible to retrieve email databases and extract relevant evidence from the e-mails;
9. *Digital timeline creation*: it shall be possible to rebuild the sequence of events and actions happened in a specific timeframe;
10. *Encrypted file retrieval*: it shall be possible to extract encrypted files;
11. *Content-based indexing*: it shall be possible to index image content to optimize queries;
12. *Cryptographic hash calculation*: it shall be possible to calculate or verify files' digest;
13. *Keyword searching*: it shall be possible to search for keywords.

2.3 Server-side functional requirements

The present section deals with non-functional requirements that complete the proposal:

- *Platform-independence*. The outlined requirements should be met regardless of the employed machine architecture and operating system.
- *Scalability*. The platform should be able to scale horizontally and the addition of more distributed machines should lead to proportional improvement in forensic tasks execution time.
- *Efficiency*. The extra work performed by the platform to distribute data and queries among its nodes, as well as to collect results, should be negligible compared to the total execution time.
- *Robustness*. Being a distributed system, the service delivery platform includes many components that potentially fail at any times. It should fall on the platform to detect and recover from such exceptional conditions and ensure the same level of confidence in the end result as in the traditional case.
- *Extensibility*. It should be easy to add a new function, as a building block of the Forensics-as-a-Service delivery platform, or replace an existing one. Therefore, writing a processing function from scratch to meet a new requirement should be compliant to the proposed model.
- *Interactivity*. To improve the user interactive experience, in such distributed solution, it should be possible to perform the time-consuming processing in the background (on cloud machines) while allowing operators to issue queries and to view partial results as soon as they become available.

- *Ease of administration.* It should be easy for administrators to operate the distributed system and minimal assumptions should be made about the underlying infrastructure (e.g., operating system services).

3. The Client-Side

In this section the author describes a Java implementation of the client-side, developed for the purpose, which integrates with the cloud-based server-side described in section II and implements the requirements discussed later on. The author used some popular forensic Linux-based distributions (e.g., Helix, Knoppix and DEFT Linux) as a reference point for reusing well-known extraction tools. The client-side of the proposed platform, running on a mobile device (e.g., a smartphone or a tablet) with Android operating system, is thought to be portable and very flexible. It is in charge of acquiring, dividing into independent fragments of smaller size and uploading logical and physical images of digital artefacts, in parallel, to the server-side platform, and sending consequently processing requests, based on the acquired data, to the server-side.

3.1 Client-side architecture

The client-side outlined above, in turn, is logically divided into two parts: a target-side, including a set of Windows extraction tools, installed on a data partition of the micro SD card, and an app-side, including an Android App installed on the OS partition, which includes the graphical user interface and the client software. The bidirectional interaction between the app-side and the target-side is straightforward as it is possible to use the micro SD card as a repository that can be used by both the app-side and the target-side. The app-side, indeed, is in charge of configuring (e.g., writing) script files on the micro SD and reading the data acquired from the target-side. The target-side, in turn, is responsible for reading the configuration scripts, executing them, collecting and writing results on the micro SD. The implemented solution avoids, therefore, complex protocol interactions and synchronization between the parties. The client-side architecture, including the described functional components, is summarized in Figure 2.

3.2 Client-side functional requirements

Client-side implementation requirements, with regards to the app-side, have been discussed in the previous section. To summarize, the app-side shall be able to:

- Manage the graphical user interface which interacts with the target-side for live and post-mortem artefact analysis;
- Configure scripts on the micro SD that launch forensic acquisition tools on the target-side;
- Read results written by the target-side on the micro SD and show them via graphical user interface;
- Redirect the target-side acquisition towards an external storage device.

Most of the requirements of the target-side are based on the consideration that a live analysis of powered-on artefacts may be critical during an investigation as it may provide important clues to first responders on the crime scene. In particular, powered-on artefacts contains memory-resident information, such as passwords and encryption keys, that will be lost after rebooting. It also happens that artefacts under

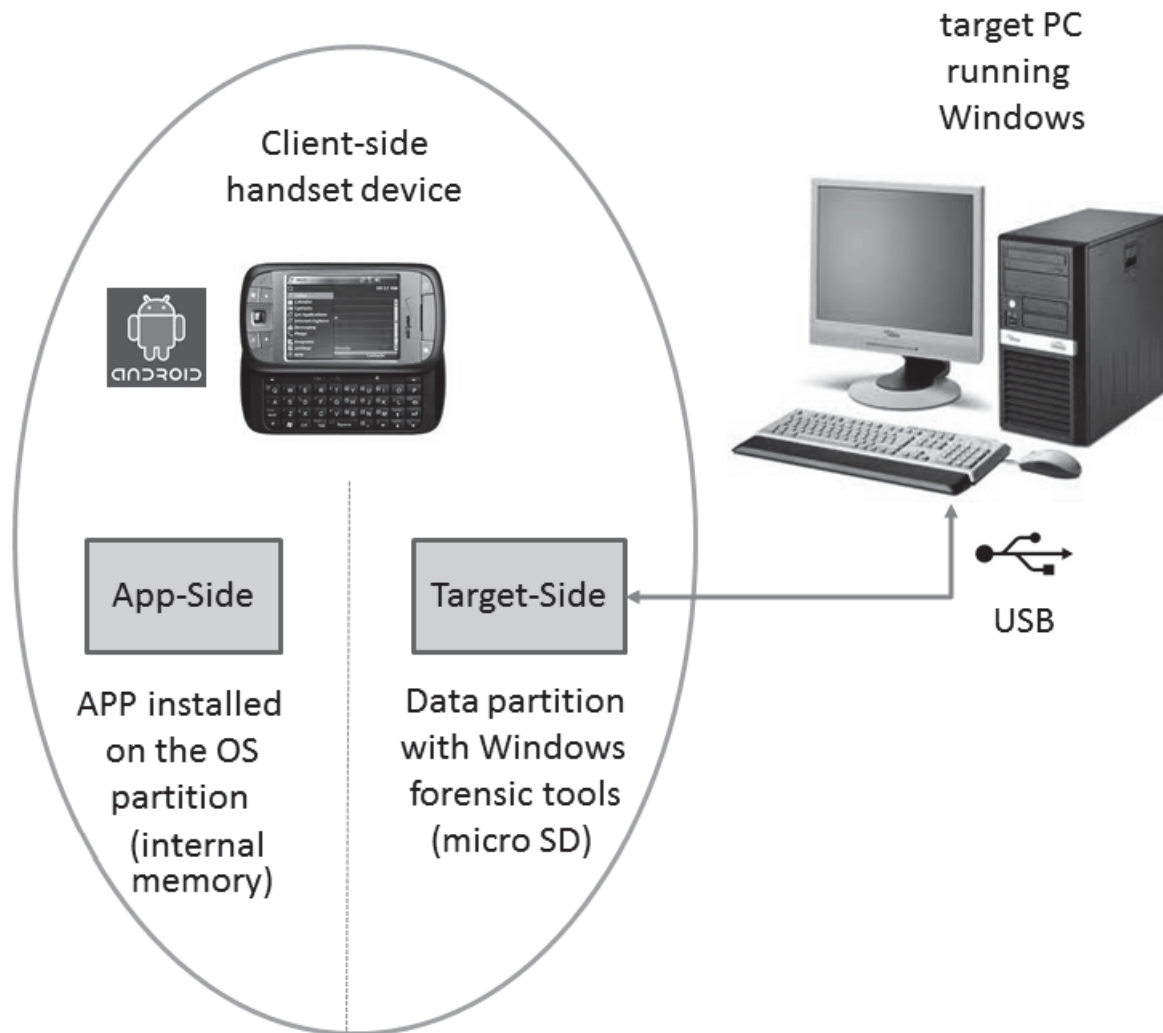


Figure 2. Client-side architecture

investigation, such as critical servers, may not be powered-off or afford disruption to service, and must be analysed with live forensic techniques. In a live forensic scenario, the target-side shall be able to:

- Acquire current system time against an accurate time source;
- Acquire and analyse artefact data in order of volatility (OOV):
 - Physical memory dump, open files, open network connections, swap space;
 - Encrypted file systems where you do not have key to unlock;
 - List of active processes;
 - Windows registry;
 - Temporary file systems;
 - Message digests of gathered evidence.

In a post mortem forensic scenario, the target-side shall be able to:

- Acquire a logical or physical image of each powered-off artefact;
- Analyse local file systems to recover specific files;
- Recover deleted files from unallocated space and slack space;

- Analyse windows registry to find installed applications, connected USB devices etc.;
- Create a timeline of events;
- Search for hidden data (steganalysis).

Part of such requirements have been already implemented in the current release of the client-side. Details about planned future implementations are provided in the next sections.

4. Conclusion

A traditional digital investigation implies that, using stand-alone forensic workstations, analysts are able to perform various forensic tasks in sequence, against limited datasets extracted from target artefacts, and evaluate correspondent results. Unfortunately this scenario is changing rapidly as the growing size of storage devices, on the one hand, and proliferation of multi-vendor mobile devices, on the other hand, are causing investigation delays and increasing complexity of digital forensics tasks. As a consequence, there is an urgent need to find novel solutions to improve digital investigation effectiveness. To solve the issues outlined above, the paper described an investigative platform for distributed forensic data processing, aimed at taking advantage of both mobility and cloud computing capabilities, called “forensics-as-a-service delivery platform.” The platform consists of a server-side spread across a multitude of commodity servers in the cloud, and a client-side running on a mobile device (e.g., a smartphone or a tablet) with Android operating system, which can be easily deployed on the crime scene to perform live and post mortem forensic tasks. The proposed solution is based on the assumption that each forensic task may be divided into a set of independent subtasks, running in parallel on a multitude of commodity servers in the cloud. It is possible to identify forensic images as the natural units of distribution among the cloud servers since independent file operations, such as keyword searching, may be split and performed in parallel on different servers. In the distributed scenario where few dependencies place constraints on the distribution of files or file fragments, it may be quite easy to reduce delays and increase efficiency by caching each subtasks’ dataset in remote servers’ RAM, as long as the subtask is up and running. The proposed cloud-based platform will take the responsibility of activating remote processing tasks and collecting partial results, allowing complex operations to be quickly performed. The client-side of the proposed platform is very important as well, as it may provide investigators with a wealth of forensic processing capabilities on the crime scene, besides being the access point to the forensic cloud. Architecture components, interfaces, functional and non-functional requirements of both server-side and client-side have been examined to provide the reader with interesting implementation guidelines. A prototype implementation of the client-side has been described in some detail.

5. Future work

Given the relevance of the topic, a growing interest is emerging among forensic practitioners towards cloud computing implications to digital forensics. Within few years, the demand for processing digital tasks “as-a-service” will probably increase among practitioners and investigators who will experience this new way of performing forensic investigations. Possible future research areas on the subject may be:

- Extending the client-side capabilities to implement: (a) a secure platform for on-demand remote support on the crime scene by lab analysts, (b) live windows registry acquisition and analysis, (c)

live windows temporary and log file acquisition and analysis, (d) live and selective file system preview and analysis and (e) live extraction of mounted encrypted partitions content.

- Performing and evaluating benchmarks to compare advantages and drawbacks of delivering forensic support services in house or contract out to a specialized cloud service provider, from the technical, economic, organizational and legal standpoint,
- Performing and evaluating benchmark with traditional digital forensics tools and techniques that may be adopted in similar situations.

Acquiring OS X File Handles Through Forensic Memory Analysis

Andrew F. Hay¹ and Gilbert L. Peterson²

Center for Cyberspace Research, Air Force Institute of Technology

¹*ahay@ieee.org*; ²*gilbert.peterson@afit.edu*

Abstract:

Memory analysis has become a critical capability in digital forensics because it provides insight into system state that cannot be fully represented through traditional media analysis. The volafox open source project has begun the work of structured memory analysis for OS X with support for a limited set of kernel structures. This paper addresses one memory analysis deficiency on OS X with the introduction of a new volafox module for parsing file handles associated with running processes. The developed module outputs information comparable to the UNIX `lsOf` (list open files) command, which is used to validate the results.

Authors:

Andrew Hay received a M.S. from the Air Force Institute of Technology and completed a thesis based on the research in this paper. He has a B.S. in Computer Science from the University of Arizona. Research interests include mobile device forensics, memory analysis, and network security.

Gilbert L. Peterson is an Associate Professor of Computer Science at the Air Force Institute of Technology, and Vice-Chair of the IFIP Working Group 11.9 Digital Forensics. Dr. Peterson received a BS degree in Architecture, and an MS and Ph.D in Computer Science at the University of Texas at Arlington. He teaches and conducts research in digital forensics, and statistical machine learning.

1. Introduction

This paper describes implementation of a new forensic capability for parsing open file information from OS X memory captures. When open files are mapped to a process, the forensic examiner learns which resources the process is accessing on disk. This listing is useful in determining what information may have been the target for exfiltration or modification on a compromised system. File handles may also help identify a suspicious process when unexpected file access or modifications are observed. Carvey further describes how a list of open files can compliment disk analysis to identify files of interest during an investigation (2009 132). Because open files can help characterize process activity and highlight misuse of a computer, it is highly desirable to recover this information from memory.

To support the extraction of file handles from raw memory two research objectives are defined. These include the design recovery of kernel data structures responsible for handling open files, and development of a flexible process for programmatically handling structures defined for different kernel architectures and operating system (OS) versions. This necessitates extensible software design resilient to changes in future versions of OS X.

Project volafox offers an open source memory analysis solution for OS X and FreeBSD written in Python (Lee 2011). Revision 52 of the source code has support for a limited set of kernel structures to parse hardware information, system build number, process listing, socket connections, loaded kernel extensions, and the syscall table. The project is extended by this research with the design recovery, structure template process, and new volafox module for parsing file handles presented in Section 3.

To validate the handles module, the UNIX command line tool `lsuf` (list open files) is used to baseline the state of open files for comparison. Testing was accomplished in a controlled environment using four virtual installations of OS X across two OS versions (10.6 Snow Leopard, 10.7 Lion) and two kernel architectures (i386, x86_64). Output from the `lsuf` command is compared with the handles module using an automated script that classifies differences according to a taxonomy in order to filter the results for further analysis.

2. Background

The memory forensics process consists of two parts. First, a copy of the target's memory called an *image* is written to external media. This requires a toolkit consisting of software to perform the capture and a USB device or network connection to preserve the image. Second, the image is analysed on a forensic workstation using tools to extract human-interpretable information.

2.1. Mac Memory Acquisition

Two imaging methods for OS X are considered in this research. First, the memory backup file saved by the host system of an OS X virtual machine (VM) during suspension can be copied to image the guest memory (Ligh et al. 2011, 577). However, due to tight hardware-software integration on the Mac it would be rare to encounter such an installation in the field, thereby limiting its usefulness to the forensic investigator.

Suiche (2010) demonstrates a second method using emulation of `/dev/mem` to dump RAM after retrieving critical symbols needed to build a kernel memory manager. This capability is available as an OS X kernel extension with the Mac Memory Reader tool (Architecture Technology Corporation [ATC] 2011). There are several disadvantages to this form of acquisition. First, loading the kernel module needed to browse full memory address space requires administrator privileges. Second, output from such a tool could be corrupted by the presence of memory forensic countermeasures (Haruyama and Suzuki 2012) or advent of a rootkit explicitly designed to subvert collection. "Fortunately, unless the subversion mechanism is very deeply embedded in the OS, a substantial amount of overhead may be incurred to prevent acquisition, potentially revealing the presence of a malicious agent" (Case et al. 2008, 2). Finally, because software must be executed on the target to perform the capture, its use alters system state. Despite the disadvantages, availability of this robust acquisition capability for the Mac encourages additional research and emphasis on analytic capabilities for the platform.

2.2. Project Volafox

This file handle parsing research adds a module to the existing open source volafox project (Lee 2011). Figure1 shows a summary of the source files from the volafox package relevant to OS X memory analysis with public classes in bold. Connections represent file dependencies, which are labeled with the public function names. The new open files module (`lsuf.py`) is shown but not discussed until Section 3.

The project `main()` and `volafox` class found in `volafox.py` are responsible for marshaling the remaining files and classes to perform memory analysis. This source file interfaces with the new file handles module (`lsuf.py`), and a number of other files indirectly related to its functionality.

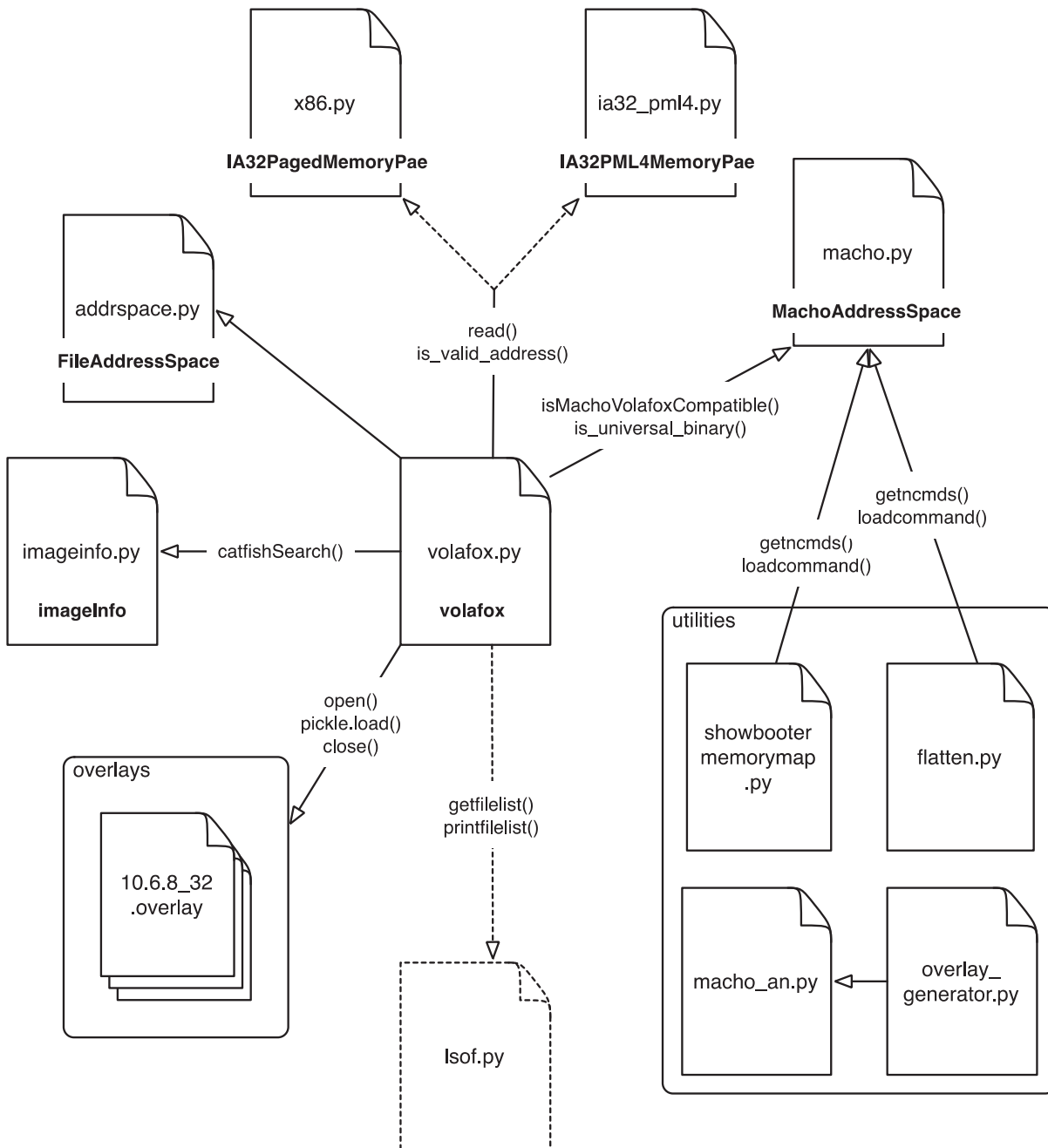


Figure 1. Volafox release 52 package diagram.

The `x86.py` and `ia32_pml4.py` files house the address space agnostic classes `IA32PagedMemoryPae` and `IA32PML4MemoryPae` respectively. They are responsible for performing virtual to physical address translations that can subsequently be converted to file offsets by `FileAddressSpace`. All requests for reading raw memory are passed through one of these two objects. PML4 is a reference to the 4th level page map used by the Intel IA-32e paging scheme (Intel

Corporation 2012), meaning `ia32_pm14.py` is responsible for handling 64-bit images where `x86.py` is used for 32-bit.

Kernel architecture is determined using the `imageInfo` class, which also returns the OS build version. This number is needed to select the correct overlay file containing the symbol list for a particular version of OS X. All symbols are read from files in the `overlays` directory using the Python `pickle` library for object serialization. New overlays can be generated from the kernel executable (`mach_kernel`) using the `overlay_generator.py` utility.

Revision 52 of `volafox`, the version extended for this research, does not natively support the Mac Memory Reader (MMR) output format. Leat (2011) and ATC developer Hajime Inoue contributed experimental support for MMR which is operational in revisions 23-38 on the project website. The feature was later removed with the introduction of 64-bit addressing support due to compatibility problems. A stand-alone `flatten.py` utility authored by Inoue is still available to convert MMR files to a linear format, but only works for 32-bit kernel installations. This utility was employed to analyse the real-world memory captures discussed in Section 4.3.

2.3. Structured Memory Analysis

Most references to memory analysis on OS X discuss context-free techniques such as string searches, manual hex examination, and file carving (Valenzuela 2011; Malard 2011; Makinen 2008). In order to perform meaningful analysis of a memory image, an understanding of the composition and location of key kernel structures is required. Because Darwin (Apple's UNIX core for OS X) is open source, the composition of kernel structures can be determined from the header files they are defined in.

Locating the structures in memory requires a mapping of identifiers and offsets, or a kernel symbol table. Suiche notes “[s]ymbols are a key element of volatile memory forensics without them an advanced analysis is impossible” (2010). The KPCR structure can be used in Windows to get the symbols directly from memory (Dolan-Gavitt 2008), however in OS X the equivalent “kernel sections are destroyed as soon as the kernel (`mach_kernel`) is loaded” [slides] (Suiche 2010). Volatility solves this problem using a database of overlay files containing the requisite symbol tables for select Linux distributions and kernel versions (Case 2011), similar functionality is provided for `volafox` by Leat (2011). In both Linux and OS X therefore, the “easiest way to retrieve kernel symbols is to extract them from the kernel executable of the hard-drive” (Suiche 2010, 4).

Figure 2 shows key features of the `mach_kernel` executable, located at the root directory of the OS X file system. Suiche (2010) describes how knowledge of the file's structure can be used to build a symbol table for a particular build of OS X. Figure 3 demonstrates how the symbol table derived from the `mach_kernel` executable is used to parse a list of running processes. Symbol `_kernproc` provides a static address for the head of the process list, `kernel_task` (PID 0), which is unique in its use of static data structures (Singh 2006, 293). The process ID, parent's PID, command name and pointer to `struct pgrp` are members of `struct proc`. Associated username information is located in `struct session`, which is referenced by `struct pgrp`. The substructure `p_list` provides a linked-list that can be walked to parse the entire running process list.

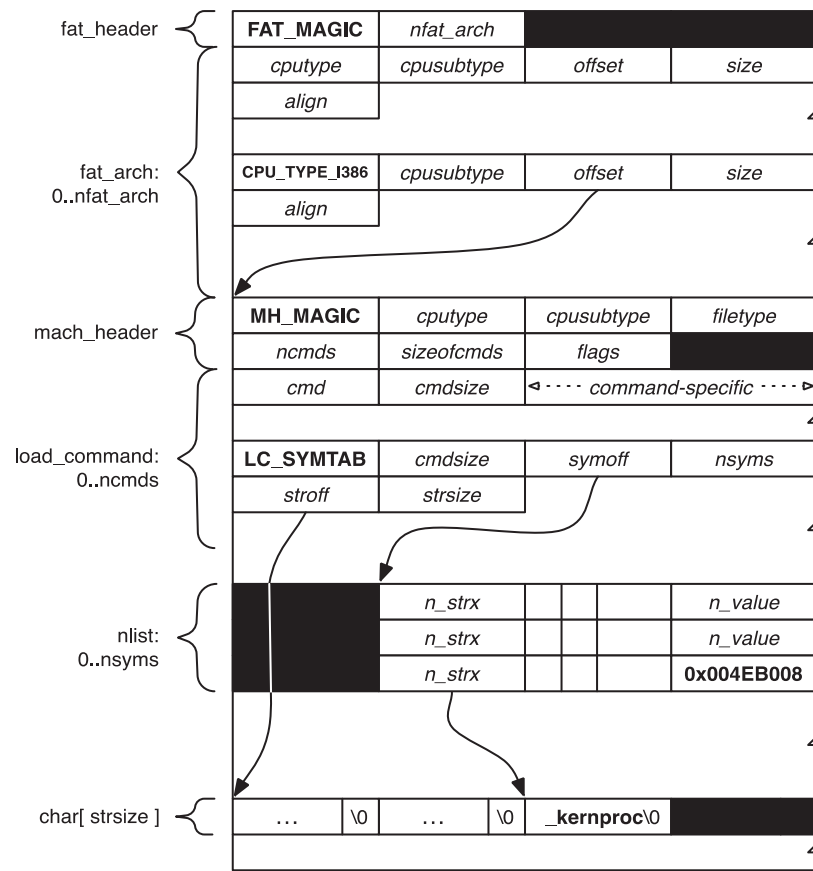


Figure 2. mach_kernel executable.

3. Methodology

The desired process-to-file handle information is an approximation of output from the UNIX `lsOf` command for OS X (Apple Inc. 2011). Because `lsOf` is included with operating system, it offers a reliable source of information for comparison. Emulating the output of this tool provides validation and offers the examiner a familiar interface to interact with. Figure 4 shows sample `lsOf` output and Table 1 describes the information in each column. The new `volafox` module for listing file handles includes functionality for parsing the nine default `lsOf` fields and the mode identifier integrated with the FD column.

While an ideal implementation would fully duplicate functionality of the `lsOf` command, due to the diversity of data structures involved the problem must be scoped for this research. Two design decisions bounding the implementation are the subset of handle types supported, and the subset of filesystems supported.

The `volafox` open files module supports handle types that subscribe to the virtual node (vnode) interface. Excluded types include POSIX semaphores and shared memory files, kernel event queue files, pipes, and sockets. These types are reported as part of the file descriptor table, but with `DEVICE`, `SIZE/OFF`, `NODE` and `NAME` fields unsupported. Additionally, the UNIX `lsOf` command classifies sockets by a variety of subtypes, which the `volafox` open files module groups together using the generic type description 'SOCKET'.

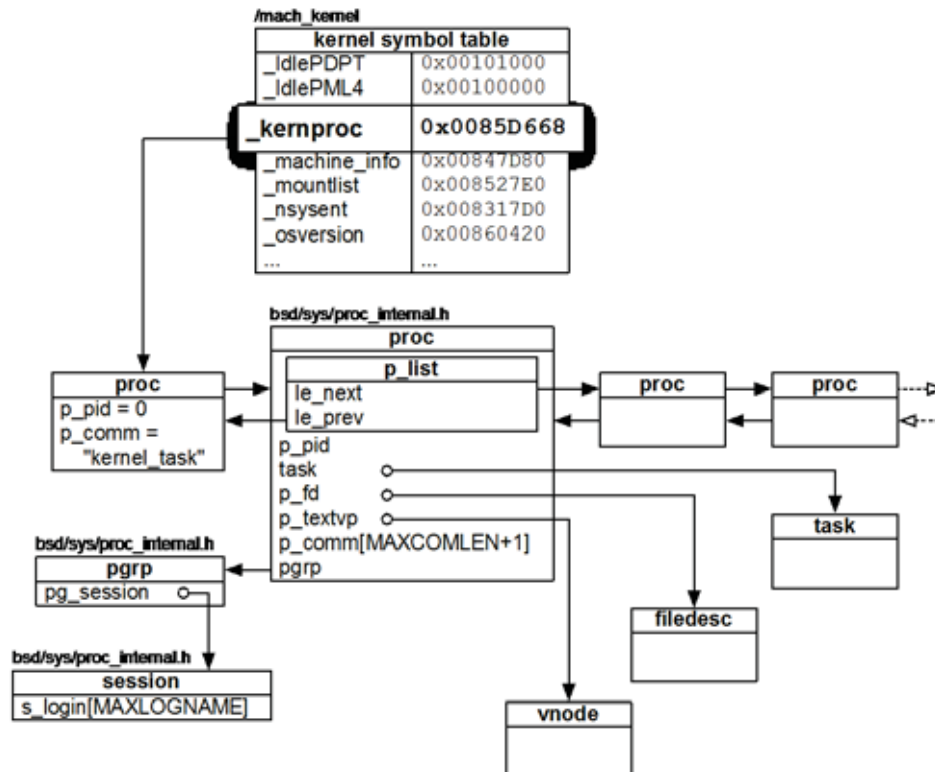


Figure 3. Symbol table and process list.

```

$ lsof -p 109
COMMAND  PID  USER  FD  TYPE  DEVICE  SIZE/OFF  NODE  NAME
bash     109  6ad   cwd  DIR   14,2    578  202041 /Users/6ad
bash     109  6ad   txt  REG   14,2   1346544  262558 /bin/bash
bash     109  6ad   txt  REG   14,2   1054960  264388 /usr/lib/dyld
bash     109  6ad   txt  REG   14,2  213385216  466405 /private/var/db/
                                dyld/dyld_shared
                                _cache_x86_64
bash     109  6ad    0u  CHR   16,0      0t369   611 /dev/ttys000
bash     109  6ad    1u  CHR   16,0      0t369   611 /dev/ttys000
bash     109  6ad    2u  CHR   16,0      0t369   611 /dev/ttys000
bash     109  6ad  255u  CHR   16,0      0t369   611 /dev/ttys000

```

Figure 4. UNIX list open files (lsof) command.

Filesystem support includes HFS+ and DEVFS. HFS+ is the default format for the OS X boot volume and DEVFS is used to abstract certain devices such as special character files. Among other uses, special character files describe ttys devices controlling the print streams stdin, stdout, and stderr for terminal programs. HFS+ and DEVFS account for the filesystems most commonly encountered during development and testing, but the vnode interface makes reference to at least 20 other types. One impact of this constraint is that files stored on network filesystems, FAT32, NTFS and others, do not have volafox support for lsof fields outside the vnode interface.

Table 1. UNIX `lsOf` output fields.

COMMAND	First nine characters of the UNIX command associated with the process.
PID	Process identification number.
USER	Login name of the user to whom the process belongs.
FD	File descriptor is a numeral index into the process open handle array optionally followed by a mode identifier: <code>r</code> (read access), <code>w</code> (write access), or <code>u</code> (both). Two other descriptors commonly seen are <code>cwd</code> representing the current working directory for the process and <code>txt</code> used for program text (code and data). These files are of high forensic value because they include the executable from which the command was launched, linked libraries, and other memory-mapped files. See the output section of the <code>lsOf</code> manpage for a full list of descriptors used (Apple Inc., 2011).
TYPE	Node type associated with the handle. See the output section of the <code>lsOf</code> manpage a partial type listing (Apple Inc., 2011). Note that numerous undocumented types were encountered in testing such as <code>FSEVENT</code> .
DEVICE	Major and minor device numbers separated by a comma. The first number describes a class of hard/software device and the second is a unique identifier for a particular instance of that class.
SIZE/OFF	Size or offset of a file reported in bytes. Offsets are preceded by a leading <code>0t</code> to distinguish when the column is mixed.
NODE	The node number for a local file. This unique identifier is filesystem dependent. For example, files stored on HFS+ report the catalog node identifier (CNID) for this field, whereas DEVFS files use a UNIX inode number instead.
NAME	Mount point and file system on which a file resides, or name of character special device.

3.2. Kernel Design Recovery

Developing the new `volafox` module for listing open files requires an understanding of 32 unique C data structures from the OS X source code, four of which are described by Suiche (2010) to list running processes. These include 26 structure (`struct`), three enumeration (`enum`), and three union definitions. Identifying the data structures containing critical information and the relationships between them is one of the primary contributions of this research because “the kernel isn’t heavily commented and its internals aren’t documented, so you learn by tracing code by hand” (Sesek 2012). Figure 5 shows an overview of the relevant structures and `lsOf` fields, associated with a particular handle type in the subscript when necessary. Using this map, the method for parsing the process list is expanded to include process handles. The structure and member associated with each `lsOf` field is shown in Table 2, however a few additional details are needed to understand the linked data structures and data decoding.

Structure `task`, as pointed to by `proc` in Figure 6, provides a link to program text files (FD `txt`). Note that each memory object may reference a `struct vm_object` or recursively refer to another entry. Memory mapped files are backed by a `vnode pager`, but the pager may be located in the shadow object for external memory managers (Singh 2006, 571).

The file descriptor table and current working directory are referenced from `struct filedesc` as shown in Figure 7. Member `filedesc.fd_ofiles` is a pointer to the start of a `fileproc` array. Elements of the array that contain a valid `fileglob` pointer reference a handle, those that do not are available to hold one. The array index represents the numerical file identifier used by the FD field of the `lsOf` output. The integer `filedesc.fd_lastfile` indexes the last file in the array and provides an iteration bound. The array itself makes up the file descriptor table, used by a process to reference all open files (ASCII, word processing, logs, temp, etc.). The file mode, also known as read/write access, is determined from the value of `fileglob.fg_flag` using the bitmap definitions in `bsd/sys/fcntl.h`.



	! DTYPE _VNODE	DTYPE_VNODE				
		VT_HFS				VT_DEVFS
		VREG	VDIR	VLNK	VFIFO	
COMMAND	proc.p_comm					
PID	proc.p_pid					
USER	session.s_login					
FD & mode	filedesc.fdesc_ofiles[i] + fileglob.fg_flag					
TYPE	fileglob. fg_type	vnode.v_type				
DEVICE		mount.vfsstatfs.fsid.val[0]				specinfo. si_rdev
SIZE/ OFF		ubc_info. ui_size	cnode. cat_attr. cau_entries	filefork. cat_fork. cf_size	fileglob. fg_offset	
NODE		cnode.cat_desc.cd_cnid				devnode. dn_ino
NAME		mount.vfsstatfs.f_mntfromname + recurse(vnode.v_parent->v_name) + vnode.v_name				

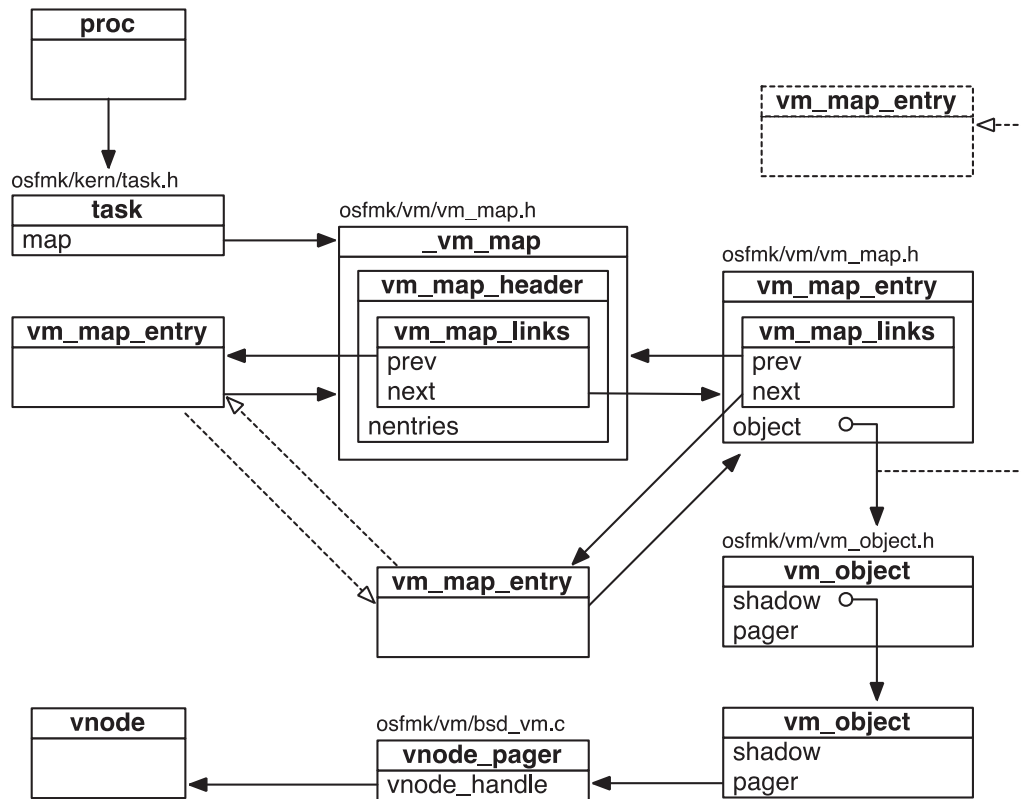


Figure 6. Memory-mapped files (txt).

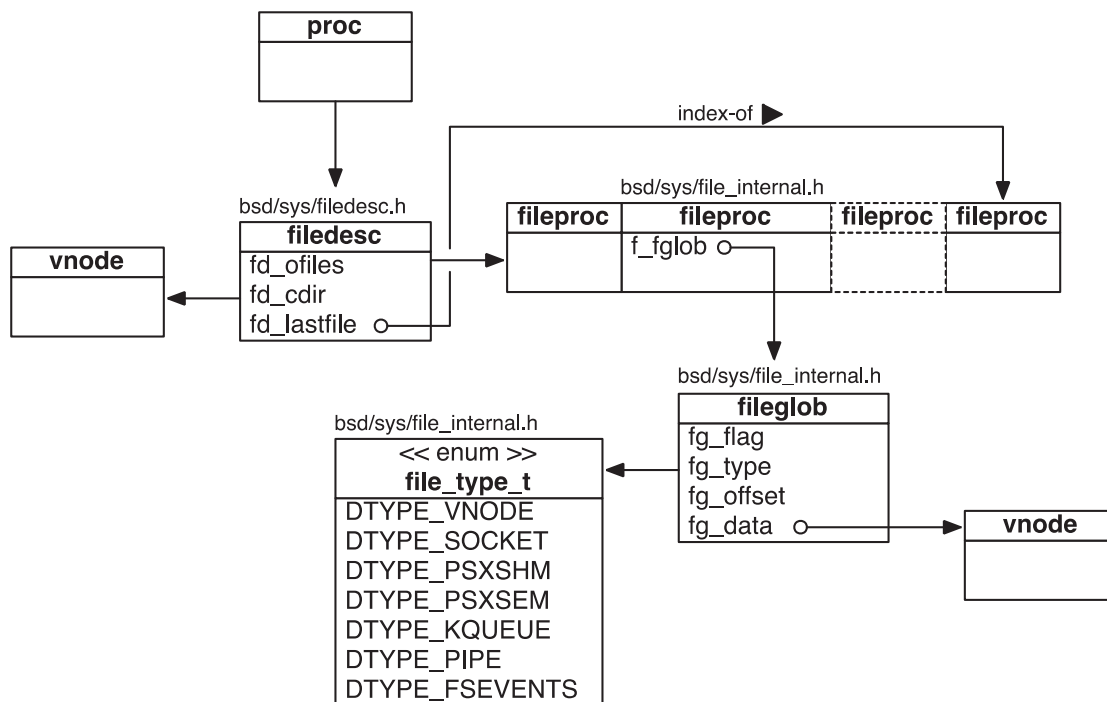


Figure 7. File descriptor table.

Determining the handle device involves additional logic because the values are encoded. The device identifier parsed from either `struct mount` or `struct specinfo` (depending on the handle filesystem) is decoded using macros in `bsd/sys/types.h` to return major and minor device number. Similarly, returning the correct SIZE/OFF value for directories requires calculation using a count value found in a substructure of `struct cnode`, the equation:

$$(entries + 2) * AVERAGE_HFSDIRENTRY_SIZE$$

found in `bsd/hfs/hfs_vnops.c`, and the macro definition from `bsd/hfs/hfs.h`.

3.3. Structure Templates

Existing volafox modules are not readily extensible and require additional logic branching for each variant in size or composition of the underlying kernel data structures. This section documents a solution developed for flexible analysis of multiple kernel architecture and OS versions. The solution consists of two parts. First, an interface is specified to describe data structures called templates. Next, a dynamic mechanism using software classes is described for selecting the correct template for a particular memory image based kernel architecture and OS version at runtime. A method for automated generation of the templates is also presented.

3.3.1 Template Interface

The following interface is defined for required members of each structure:

```
template = { MBR_NAME : ( MBR_TYPE, OFFSET, SIZE, FIELD,
                          SUB_STRUCT ), ... }
```

Table 3 lists Python types from the kernel structure template interface, which itself is implemented as a dictionary. Substructures are defined as those contained within the memory allocated for a super structure. They share the same dictionary format as regular structures and their values are referenced recursively. Figure 8 shows the 32-bit Snow Leopard variant of the `struct proc` template as an example. To support the test cases described in Section 4.2, three additional templates for the process structure are defined, one for each combination of OS version and architecture.

Table 3. Template interface fields.

Variable	Python Type	Description
template	dict	template implementing the C struct interface
MBR_NAME	str	dictionary key, variable name for a struct member
template[MBR_NAME]	tuple	dictionary value, a struct member description
MBR_TYPE	str	C type of the named member
OFFSET	int	offset in bytes for the member
SIZE	int	size in bytes for the member type
FIELD	str	ls of field represented by member
SUB_STRUCT	dict	recursively defined substructure (<i>optional</i>)

```

template = {
    'p_list' : ( 'LIST_ENTRY(proc)', 0, 8, '' , {
        'le_next' : ( 'struct proc *', 0, 4, '' ),
        'le_prev' : ( 'struct proc **', 4, 4, '' )
    }
),
    'p_pid' : ( 'pid_t', 8, 4, 'PID' ),
    'task' : ( 'void *', 12, 4, '' ),
    'p_fd' : ( 'struct filedesc *', 104, 4, '' ),
        'p_textvp' : ( 'struct vnode *', 388, 4, '' ),
    'p_comm' : ( 'char[]', 420, 17, 'COMMAND' ),
    'p_pgrp' : ( 'struct pgrp *', 472, 4, '' )
}

```

Figure 8. struct proc template, for Mac OS X 10.6 on x86 hardware.

3.3.2 Template Selection

The second component in the template solution is a Python class initializer that dynamically selects the correct template for a given subclass based on the OS version and architecture of the memory image under analysis. Because classes in the open files module manage fields and methods associated with a particular kernel structure, all inherit from the abstract superclass in Figure 9.

```

class Struct(object):

    mem      = None
    ver      = False
    arch     = -1
    kvers    = -1

    TEMPLATES = None
    template  = None
    ssize     = -1

    def __init__(self, addr):

        if self.__class__.template == None:

            self.__class__.template = self.__class__.TEMPLATES[Struct.arch] \
                                     [Struct.kvers]

            for item in self.__class__.template.values():
                if ( item[1] + item[2] ) > self.__class__.ssize:
                    self.__class__.ssize = item[1] + item[2]

        self.smem = Struct.mem.read(addr, self.__class__.ssize);

```

Figure 9. Simplified abstract class Struct (no error handling).

The first four static variables belong to the abstract class and are shared by all Struct subclasses. The mem variable is a reference to one of the PAE objects responsible for virtual-to- physical address

translation. Verbose flag `ver` indicates if *all* file descriptors should be printed, including those for types not fully supported by the open files module. The `arch` and `kvers` variables report the kernel architecture and version respectively. The final three fields are virtual static variables because their assignment is deferred to the subclasses. The constant `TEMPLATES` is a nested dictionary from which the static `template` is assigned the first time the initializer runs based on value of `arch` and `kvers`. The static `ssize` is subsequently assigned based on the selected template and determines how many bytes the initializer reads from the address passed as an argument to provide coverage of all members specified in the structure template.

Combining the structure template interface with an abstract initializer offers a solution that greatly simplifies program logic needed to support a selection of architectures and OS versions. The result is also highly extensible because new templates can be added without any code refactoring as long as the member names remain consistent across versions. Figure 10 shows the concrete subclass corresponding to `struct devnode` and demonstrates use of the structure template solution.

```
class Devnode(Struct):

    TEMPLATES = {
        32:{
            10:{'dn_ino':('ino_t',112,4,'NODE')}
            , 11:{'dn_ino':('ino_t',112,4,'NODE')}
        },
        64:{
            10:{'dn_ino':('ino_t',192,8,'NODE')}
            , 11:{'dn_ino':('ino_t',192,8,'NODE')}
        }
    }

    def __init__(self, addr):
        super(Devnode, self).__init__(addr)

    def getnode(self):
        return unpacktype(self.smem, self.template['dn_ino'], INT)
```

Figure 10. Concrete class Devnode.

3.3.3 Member Offsets and Type Sizing

While the dictionary constants used to implement structure templates are easy to work with programmatically, generating their syntax is labor intensive. The new open files module uses (*18 classes * 2 versions * 2 architectures*) = 72 structure templates, requiring a great deal of error-prone coding and debugging if generated by hand. Determining *size* and *offset* values for each member in the template is also very difficult to accomplish manually due to the complexity of defined types included in the kernel structures. The solution to both of these challenges is an external C program that dissects kernel structures and automates the generation of the Python dictionary syntax needed for each template.

The `offsets.c` program was developed to find the size and offset of each required structure member and print the results as a structure template for use in `lsOf.py`. Figure 11 shows a function from the program that prints a template for `struct _vm_map`. The variable member is a C structure

```

int vm_map() {
    member m;
    int (*mh)(unsigned long int offset) = &vm_map_header;

    printf("struct_vmmmap = {");

    m.var_name = "hdr";
    m.var_type = "struct vm_map_header";
    m.offset = offsetof(struct _vm_map, hdr);
    m.size = sizeof(struct vm_map_header);
    m.field = "";
    printmember(m, mh);

    printf("}\n");
    return 0;
}

```

Figure 11. Template generation function.

defined in the program to hold the fields described in Table 3 and `printmember()` formats each as a key/value pair for the enclosing Python dictionary. The argument `mh` is a function pointer to a substructure that is printed recursively.

The C language `sizeof` operator is used to find the size of any type, and the preprocessing macro `offsetof` defined in `stddef.h` can return the offset of any member for a given structure. However, most of the header files defining kernel structures are not available in the include path for OS X. Sesek (2012) explains the problem and suggests a workaround in a blog post about kernel debugging:

Structs [...] are merely human-friendly offsets into a region of memory. Their definition and layout can be shamelessly copied from the XNU open source headers into your kext's project so that you can access fields in kernel private structures. As it turns out, virtually ever structure within the kernel is designed to be opaque to a kext. Apple decided to do this so that they can freely change the kernel structures, but it also makes writing a debugging tool like this a little harder. To do so you need to edit the headers so they compile in your project through a process I call "munging."

Sesek's method was modified to access the kernel definitions needed for `offsets.c` using local headers.

Three out of 18 template functions written for `offsets.c` are known to produce incorrect member offsets for 64-bit kernel architecture. The problem is believed to be a complex definition conflict for some low-level types. Several C types are defined for userspace with standard libraries such as `stdio.h`. However, the kernel sometime uses different sizes for these same types and forced redefinition yields a compilation error. When the `offsetof` macro measures a userspace definition the result is an error for some architectures. Manual offset calculation and hex analysis are used to resolve the problem for affected templates, resulting in adjustment made to the `TEMPLATES` constant of the equivalent structure class in `lsof.py`.

A wrapper for `offsets.c` called `printstructs.py` is written to verify the output dictionary as executable Python code, print the structure members in a human-readable format for debugging, and

handle compilation flags related to architecture. Dictionary output from `printstructs.py` was then pasted into the `TEMPLATES` constant of each class in `lsof.py` to complete the definition.

3.4. Open Issues

There are two outstanding problems with the research module developed: reporting the correct user associated with a process, and determining size of the `/dev` directory.

The manpage for the UNIX `lsof` command describes output of the `USER` field as “the user ID number or login name of the user to whom the process belongs, usually the same as reported by `ps(1)`.” However, output from the `volafox` open files module is known to incorrectly report the process login name as shown in Figures 12-13.

```
$ ./volafox.py -i 10.6.8x86.vmem -o lsof -p 15
COMMAND  PID  USER  FD      TYPE [...]
distnoted 15  root  cwd      DIR [...]
distnoted 15  root  txt      REG [...]
distnoted 15  root  txt      REG [...]
distnoted 15  root  txt      REG [...]
distnoted 15  root  0r       CHR [...]
distnoted 15  root  1        PIPE [...]
distnoted 15  root  2        PIPE [...]
distnoted 15  root  3u       KQUEUE [...]
distnoted 15  root  56u      SOCKET [...]
```

Figure 12. volafox user output.

```
# lsof -p 15
COMMAND  PID  USER [...]
distnoted 15  daemon [...]
distnoted 15  daemon [...]
distnoted 15  daemon [...]
distnoted 15  daemon [...]
distnoted 15  daemon [...]
distnoted 15  daemon [...]
distnoted 15  daemon [...]
distnoted 15  daemon [...]
distnoted 15  daemon [...]
distnoted 15  daemon [...]
```

Figure 13. lsof user output.

The difference shown is not consistent across all processes of a full file listing. In many cases the expected `USER` value is reflected in the output, but not always. This problem is not unique to the `volafox` open files implementation as it is also present in the `volafox` process listing module and the original work on which it was based (Suiche 2010). Kernel structure analysis of the source headers could not identify an issue with Suiche’s methodology, but all tests indicate that `struct session` cannot consistently return the username for any of the user-related keywords available for `ps`. There is also no known method to determine when the `session` structure returns the correct value.

A second problem identified during development is an inability to correctly report the `SIZE/OFF` field for certain directories. The `/dev` directory is typed `DTTYPE_VNODE` in `fileglob.fg_type` and `VDIR` in `vnode.v_type`. However, it has a tag of `VT_DEVFS` from `vnode.v_tag` rather than the `VT_HFS` seen for most other directories. Figure 14 shows an example of `/dev` as reported by the UNIX `lsof` command.

```
# lsof +d /dev
COMMAND PID USER  FD  TYPE      DEVICE SIZE/OFF  NODE NAME
launchd  1  root   8r   DIR 20,5853800  4495  305 /dev
```

Figure 14. `/dev` directory size.

Note that $4495 \bmod 34 \neq 0$, and therefore sizing by the entry count as described in Section 3.2 is not valid for this directory.

Table 2 includes three alternate locations for the size applicable to other file types, but none were found to be effective in this case. Fortunately, due to the unique combination of tag and type for `/dev`, the failure is possible to detect. Since the location of the size is unknown, the volafox open files module prints `-1` for the size of `/dev` to indicate the field is unsupported.

4. Testing

Testing effectiveness of the volafox module for listing OS X file handles involves comparing its output with that of `lsOf`. A successful implementation of the system must accurately report all file handles, adjusted for stated constraints and known deficiencies. Testing is conducted on controlled test cases and on captures from real user's machines.

The complex nature of a modern operating system like OS X guarantees changes to the system state between the time when the `lsOf` command is run and the memory dump occurs (Hay and Nance 2009). Some allowance is necessary to account for volatility of the handles list during this interval. A successful implementation therefore becomes one that can be validated against the UNIX `lsOf` command, adjusted for stated constraints, known deficiencies, and accuracy of the validation method.

Table 4. Field differences versus file type.

File Type	COMMAND	PID	USER	FD+ mode	TYPE	DEVICE	SIZE/ OFF	NODE	NAME
cwd	✓	✓	D1	✓	✓	✓	✓	✓	✓
txt	✓	✓	D1	✓	✓	✓	✓	✓	✓
REG	✓	✓	D1	✓	✓	✓	✓	✓	✓
DIR	✓	✓	D1	✓	✓	✓	D2	✓	✓
CHR	✓	✓	D1	✓	✓	✓	E7	✓	✓
LINK	✓	✓	D1	✓	E5	✓	✓	✓	✓
FIFO	✓	✓	D1	✓	✓	E6	✓	✓	✓
VNODE (other)	✓	✓	D1	✓	✓	C3	C3	C3	✓
PSXSHM	✓	✓	D1	✓	✓	C2	C2	C2	C2
PSXSEM	✓	✓	D1	✓	✓	C2	C2	C2	C2
KQUEUE	✓	✓	D1	✓	✓	C2	C2	C2	C2
PIPE	✓	✓	D1	✓	✓	C2	C2	C2	C2
FSEVENT	✓	✓	D1	✓	✓	C2	C2	C2	C2
SOCKET	✓	✓	D1	✓	C1	C2	C2	C2	C2

4.1. Comparison Taxonomy

A formal list of 21 differences across four categories is used to classify reasons output from `lsOf` and the new volafox handles module may differ. Taxonomic categories consist of constraints, deficiencies, explained differences, and failures. Enumeration labels are employed by the script `validate.py` to describe how similar the volafox output is to its validation data.

4.1.1 Constraints

Constraints are defined as differences in output that occur due to system design decisions. The volafox open files module has several limitations with regard to handle type and filesystem tag that are used to scope the research implementation.

- C1. The lsof subtype for socket handles cannot be determined. A value of `DTYPE_SOCKET` for the member `filglob.fg_type` indicates a socket handle. The lsof command reports a number of subtypes for these handles including: `sysm`, `unix`, `IPv4`, `IPv6`, `rte`, `key`, `ndrv`, and possibly others that were not observed in testing. Sockets are assigned the generic type `SOCKET` in the volafox open files output.
- C2. Only handles subscribing to the virtual node (vnode) interface are fully supported. A value of `DTYPE_VNODE` for the member `fileglob.fg_type` indicates the vnode interface is in use for a particular handle. Full support indicates meaningful output is reported for all nine lsof command fields. Non-vnode handles show the value `'-1'` for `DEVICE`, `SIZE/OFF`, `NODE`, and `NAME` to indicate these fields are unsupported in the volafox open files output.
- C3. Only vnodes tagged `HFS+` or `DEVFS` are fully supported. A value of `VT_HFS` or `VT_DEVFS` for the member `vnode.v_tag` indicates a supported filesystem. The lsof command fields `DEVICE`, `SIZE/OFF`, and `NODE` are defined outside struct `vnode` and therefore unsupported for other filesystems. Unsupported fields are indicated in the volafox open files output with an appropriate value from `ECODE`, a global dictionary defined for `lsof.py`.

4.1.2 Deficiencies

Deficiencies are defined as differences in output that occur due to known implementation problems. As described in Section 3.4, the volafox open files module has two open issues.

- D1. The lsof `USER` field is not correctly reported for all processes in a full file listing. This problem is not consistent across all processes and the volafox open files module is not capable of detecting its occurrence.
- D2. Size of the `/dev` directory cannot be determined. Handles with `vnode.v_type` of `VDIR` and `vnode.v_tag` of `VT_DEVFS` such as `/dev` show the value `'-1'` in the `SIZE/OFF` field.

4.1.3 Explanations

Explained differences are those in output that occur due to reproducible idiosyncrasies of the tools used for capture or validation. They are distinct from failures because the explanations are not speculative, and the differences can be detected using automation. Explanations E4, E5, and E6 are believed to be bugs in the OS X version of the `lsof` program.

- E1. The UNIX lsof command output always includes the lsof command and its associated handles, whereas a memory dump does not. For 10.7 only, the dependent process `sudo` is present in addition to lsof when executed with administrator privileges.
- E2. Memory captured using the MMR tool includes handles associated with the process `MacMemoryReader` and its dependency image, whereas output from the lsof command does not.

- E3. Data collected using `capture.py` (4.3) does not share the process `sh` because `MacMemoryReader` and `lsuf` are executed in different subprocesses.
- E4. OS X duplicates some handles in a full listing using `lsuf`. Duplication occurs at least once per listing. Figure 15 demonstrates the problem.

In all observed cases, the file descriptor `'twd'` (the per-thread working directory) identifies the duplicate, while all other fields remain the same.

\$ sudo lsuf								
COMMAND	PID	USER	FD	TYPE	DEVICE	SIZE/OFF	NODE	NAME
...								
mds	29	root	cwd	DIR	14,2	1088	2	/
mds	29	root	twd	DIR	14,2	1088	2	/
...								

Figure 15. `lsuf` handle duplication.

- E5. OS X reports the type of symbolic links as `'0012'` instead of `'LINK'` in the `lsuf` TYPE field. The keyword `'LINK'` is specified in the manpage and therefore the `volafox` handles module reports symbolic links using that label. The bug has only been observed in the 10.7 version of OS X.
- E6. OS X does not report the `lsuf` DEVICE field for FIFO type files. The manpage does not discuss the omission and the `volafox` open files module can determine the major and minor device number for FIFO special files.
- E7. Execution of the `lsuf` command causes the offset of its terminal file (`ttys`) to grow. For cases where a `ttys` file is the same used by the `lsuf` command, any offset difference is classified as E7 rather than F6.

4.1.4 Failures

Failures are defined as differences in output not already accounted for by constraints, deficiencies, or explanations that occur due to asynchronous data collection or implementation artefact. It is important to note that the fault *causing* failure is undefined by default. Analysis in Section 4.2 indicates that in most cases failure is a consequence of validation accuracy rather than an error in the `volafox` open files module implementation.

- F1. Command name mismatch (field: COMMAND). Adjusted for F2.
- F2. Missing/extra process (field: PID). Adjusted for E1, E2, and E3.
- F3. Missing/extra file descriptor (field: FD). Adjusted for F2 and E4.
- F4. File mode mismatch (field: FD). Adjusted for F3.
- F5. File type mismatch (field: TYPE). Adjusted for F3, C1 and E5.
- F6. Device mismatch (field: DEVICE). Adjusted for F3, C2, C3, and E6.
- F7. Size/offset mismatch (field: SIZE/OFF). Adjusted for F3, C2, C3, D2, and E7.
- F8. Node identifier mismatch (field: NODE). Adjusted for F3, C2 and C3.
- F9. Pathname mismatch (field: NAME). Adjusted for F3, C2.

Username mismatch is classified as D1 and therefore not listed as a failure. It is reported in the results after adjustment for F2. Reporting failures F2 and F3 also aligns the process and handle lists of each file respectively for the remaining failure tests. This means, for example, that F1 does not report command name mismatches that occur due to a missing process because F2 already accounts for it.

4.2. Controlled Test Cases

Controlled test case results are examined with the goal of identifying previously unidentified implementation problems. The majority of constraints, deficiencies, and explained differences are not considered in this analysis as the failures alone describe possible unknown faults in the tool developed. The validation method conducts software test cases that either pass or fail. Resulting failures are then addressed individually, or reclassified in the difference taxonomy. Where an explanation is provided for a failure, the discussion must be viewed as speculative because all concrete differences identified have been integrated with the analysis taxonomy.

One design goal for the module developed is to provide coverage for a breadth of OS versions and kernel architectures. These test cases are intended to demonstrate that coverage by representing both i386 and x86_64 Intel architectures over the span of minor OS X versions (10.6.0-8 and 10.7.0-3) within the current and previous releases of the operating system. All tests are performed on guest installations of OS X running as a VM. This setup offers the linear file format volafox requires in analysing 64-bit kernel memory, the contents of which are written to disk when the VM is suspended. Efforts were made to minimize OS interference with the state of open files during collection. Specific modifications include: removing network interfaces, deleting startup items, and disabling the OS X automatic file indexing process known as Spotlight. The sole installation of OS X Server also had the `servermgrd` daemon disabled to eliminate its numerous child processes on startup.

Configurations for the controlled test cases include:

- | | |
|--|---|
| 1. OS X version: 10.6.8
Darwin kernel architecture: i386
RAM installed: 1 GB | 3. OS X version: 10.7.3
Darwin kernel architecture: i386
RAM installed: 2GB |
| 2. OS X version: 10.6.0 Server
Darwin kernel architecture: x86_64
RAM installed: 1GB | 4. OS X version: 10.7.0
Darwin kernel architecture: x86_64
RAM installed: 2GB |

Table 5 summarizes results across the four controlled test cases. After accounting for constraints, deficiencies, and explained differences listed in the analysis taxonomy (not shown), this table indicates how similar the volafox open files output is to the `lsOf` approximation. Failures in the comparison are marked in red and discussed in order from top to bottom of the table.

The extra process in the volafox output (F2) for the 10.7.x cases is a daemon with the highest PID in the process list. It therefore appears to have been launched after executing `lsOf`, explaining its absence in the baseline listing in both instances.

While the username deficiency (D1) is not classified as a failure, it is listed in the table to emphasize the number of handles affected by this bug.

The additional volafox file descriptors (F3) in the 10.7.3 test case, and three of the four in the 10.7.0 case belong to `launchd`. Because the `launchd` process manages all other daemons (Singh 2006, 38), it is very active and therefore volatile. For both 10.7.x test cases the `lsOf` and `launchd` processes appear

Table 5. Difference summary for controlled test cases.

Diff	Field	10.6.8 i386	10.6.0 Sever x86_64	10.7.3 i386	10.7.0 x86_64
F1	COMMAND	0	0	0	0
F2	PID	0	0	+1	+1
D1	USER	15%	17%	38%	19%
F3	FD	0	0	+3	+4
F4	mode	0	0	0	0
F5	TYPE	0	0	0	0
F6	DEVICE	0	0	0	0
F7	SIZE/OFF	1	1	1	1
F8	NODE	2	1	1	1
F9	NAME	0	0	0	0

to be confounded, though similar problems were not observed in the 10.6.x test cases. These differences are believed to represent normal OS interference with the state of open files between the time `lsOf` is executed and the VM is frozen.

The fourth extra file descriptor (F3) in the 10.7.0 test case appears to be a malformed vnode. All members within the structure are invalid, and the file name is made up of non-ASCII characters. This case does call into question the methodology described in Section 3.2 for determining valid descriptors in the file table. Since the occurrence appears to be isolated, it is particularly difficult to debug this potential implementation failure. One possible explanation is that the handle may be an initialized but as-yet-unused vnode in the file descriptor table. Luckily, the error output is well-handled and therefore a human analyst should be able to make this determination with ease even if the tool cannot.

In all four test cases, the file size failure (F7) is for the pseudo-tty device opened by process `Terminal`. The `Terminal` application is in the process hierarchy for `lsOf`, which as explained in E7 is known to modify some `ttys` device offsets during execution. This explanation might have led to another explained difference in the taxonomy, but detection could not be easily automated for this case.

In all four test cases, the node identification failures (F8) belong files related to time zone opened by the `notifyd` process. It is unclear why the notification server makes changes to these files during `lsOf` execution and additional knowledge of OS X internals is needed to analyse this failure further. However, because the difference in node value is always observed on regular files but only those associated with time zone and this particular process, it is not believed to be an implementation fault.

Results from the four controlled test cases yield several important conclusions. First, the `volafox` open files module is functional for kernels utilizing both Intel i386 and x86_64 architectures. Second, the tool provides coverage for the OS X 10.6.x Snow Leopard and 10.7.x Lion operating systems. Third, the username deficiency (D1) results suggest that this field cannot be trusted in the `volafox` output. Finally, the low number of unexplained failures suggests the implementation is successful under the research definition.

4.3. Real-world Data Analysis

In addition to the controlled test cases, the `volafox` handles module was also tested against a set of memory collected from physical machines. The script `capture.py` was developed to automate

collection of memory using the MMR tool and a variety of incident response data, including `lsOf`, for comparison. These real-world collections are invaluable for program debugging and revealing edge cases in the handles implementation but are not well suited for validation for several reasons. First, because failures cannot be replicated it is difficult to determine if a fault is caused by implementation bug or validation accuracy. Second, the collection time required by MMR assures that output from `lsOf` is always stale when compared to the memory capture. Finally, the real-world data available does not cover the breadth of OS versions and kernel architectures.

Revision 52 of the volafox project does not support the MMR output format directly. As a result, only i386 captures are analysed with volafox after conversion to linear format using the `flatten.py` utility. Ten qualifying samples were collected from real Mac computers, eight of these running 10.6.8, one 10.7.0, and one 10.7.2.

Table 6 shows a combined summary of the real-world results. Because the hardware and software configurations vary greatly between collections, the data points represent different sample populations that cannot be aggregated to produce valid mean or standard deviation. Instead, the range of each constraint, deficiency, explained difference, and failure is reported to offer a general impression of how commonly these differences occur. A few noteworthy conclusions emerge from this analysis.

1. With up to 10% of processes (F2) and 22% of handles (F3) thrown out for comparison during alignment, `lsOf` does *not* approximate the real-world data very closely.
2. The set of non-vnode handles (sockets, pipes, semaphores, etc.) make up a significant portion of the `lsOf` results (C2). Sockets in particular are of high investigative value and should therefore be considered in future work.
3. Unsupported file systems (C3) in the real-world data were cross-referenced with the mount information also collected by the `capture.py` script to determine which types should be considered for future support. The results included one instance each of: `msdos` (FAT32 external hard drive), `cddafs` (responsible for reading audio CDs), `ntfs` (Apple Bootcamp installation of Windows), and `mtmfs` (used to implement the Mobile Time Machine feature).
4. Explained differences (E1-E7) and the `/dev` sizing deficiency (D2) do not affect a large number of processes and handles. However, their enumeration is important because it filters the number of failures that must be considered.
5. For a given handle the size/offset (F7) and node identifier (F8) information can be particularly volatile, with up to 10 and 8 percent change observed respectively.
6. Upon manual inspection of the failures, high volatility of the name field (F9) was often linked to two applications: Spotlight and the Microsoft suite. Spotlight is Apple's indexed search technology and automatically begins processing external media when mounted. Because the `capture.py` script is delivered on external media, the act of collection increases indexing activity.

The real-world data identifies a number of implementation problems that may not have been encountered otherwise. For example, the E2, E3, and E7 results include an asterisk because one of the samples experienced an interesting collection failure. The `capture.py` script and all its associated processes (Python, `sh`, `MacMemoryReader`, `image`, etc.) are all absent from the volafox output for this sample, making it clear the processes list had been truncated. Due to the high occurrence of invalid pointers observed in the real-world data and several volafox execution errors, additional exception

Table 6. Combined real-world results (10 samples).

Diff	Description	Quantity or % Per Sample	
C1	SOCKET handles cannot be subtyped	15-22%	of handles affected
C2	Non-vnode handles are not fully supported	27-40%	of handles affected
C3	Non-HFS+/DEVFS vnodes are not fully supported	0-4%	of handles affected
D1	Δ USER field	16-54%	of usernames misreported
D2	/dev directory cannot be sized	0-1	handles affected
E1	ls of process is not shared	0-1	process removed
E2	MacMemoryReader and image processes are not shared	0*-2	processes removed
E3	sh process is not shared	0*-1	process removed
E4	Duplicate handles labeled FD: 'twd'	2-5	handles removed
E5	LINK handles are mislabeled	0-3	handles affected
E6	FIFO handles do not report device identifier	0-2	handles affected
E7	ls of ttys file size is not shared	0*-13	handles affected
F1	Δ COMMAND field	0	commands differ
F2	Δ PID field	0-10%	of processes removed
F3	Δ FD field	4-22%	of handles removed
F4	Δ MODE field	0-2	modes differ
F5	Δ TYPE field	0-2	types differ
F6	Δ DEVICE field	0-2	device identifiers differ
F7	Δ SIZE/OFF field	0-10%	of sizes/offsets differ
F8	Δ NODE field	0-8%	of node identifiers differ
F9	Δ NAME field	0-3%	of names differ

handling was added to the `ls of` module to support debugging. The new code identified several cases where the underlying linked data structures were broken in the memory image. In a real investigation these occurrences might represent evidence lost. One recommendation to mitigate this problem is to assure memory capture proceeds as rapidly as possible. One factor found to affect capture speed in real-world collections is the type of external media used to store the image. Timing results recorded by `capture.py` showed a 16 Mb/s average increase in capture speed when using an external hard drive over flash storage.

5. Conclusions and Future Work

This paper presents documentation and implementation of a new capability for parsing file handles from an OS X memory capture. Initial development of the module required performing a manual design recovery of the data structures responsible for handling files for OS X. To alleviate the manual recovery process in future versions of OS X, a novel header-processing tool programmatically parses structures defined for different kernel architecture and OS versions and converts these into templates used by the file handle module.

Testing the implementation identified several areas for future work not directly related to the research goal. First, the open files module does not reliably output the correct user of a running system

process. No fault could be identified in the implementation, nor any problem with the kernel structure analysis described in prior work. Second, memory captured on physical hardware suffers from a high number of invalid pointer references, occasionally resulting in malformed linked-lists. Robust exception handling needs to be implemented throughout volafox to address this problem in a memory analysis tool.

Finally, several additional modules must be developed to establish volafox as an analysis tool suited for technical users and forensic examiners. The National Institute of Standards and Technology (NIST 2006) describes the minimum requirements for volatile collection during incident response. At present, the volafox tool includes modules for parsing several of these requirements but is still missing a list of login session, network configuration, and operating system time.

Acknowledgements

This research was supported by the U.S Air Force Cyberspace Technical Center of Excellence and the U.S. Air Force Office of Scientific Research (AFOSR/RSL). Views expressed in this paper are those of the authors and do not reflect the official policy or positions of the U.S. Air Force, U.S Department of Defense, or the U.S. Government.

Bibliography

Apple Inc. "Isof(8) Mac OS X manual page." Last modified March 21, 2011.

<http://developer.apple.com/library/mac/#documentation/Darwin/Reference/ManPages/man8/Isof.8.html>.

Architecture Technology Corporation. "Mac Memory Reader." Accessed May 24, 2011.

<http://cybermarshal.com/index.php/cyber-marshall-utilities/mac-memory-reader>.

Carvey, Harlan. *Windows forensic analysis DVD toolkit*, 2nd ed. (Burlington: Syngress, 2009).

Case, Andrew, Andrew Cristina, Lodovico Marziale, Golden Richard, and Vassil Roussev. "FACE: Automated digital evidence discovery and correlation." *Digital Investigation* 5(Supplement) (2008): S65-S75.

Case, Andrew. 2011. "Bringing Linux support to Volatility." Last modified March 1, 2011.

<http://dfsforensics.blogspot.com/2011/03/bringing-linux-support-to-volatility.html>.

Dolan-Gavitt, Brendan. "Finding kernel global variables in Windows." Last modified April 16, 2008.

<http://moyix.blogspot.com/2008/04/finding-kernel-global-variables-in.html>.

Haruyama, Takahiro, and Hiroshi Suziki. "One-byte Modifications for Breaking Memory Forensic Analysis." Slides presented at Black Hat Europe, Amsterdam, Netherlands, March 14-16, 2012. https://media.blackhat.com/bh-eu-12/Haruyama/bh-eu-12-Haruyama-Memory_Forensic-Slides.pdf.

Hay, Brian, Kara Nance, and Matt Bishop. "Live analysis: Process and challenges." *IEEE Security & Privacy* 7, no. 2 (2009): 30-37.

Intel Corporation. "Intel 64 and IA-32 Architectures Software Developer's Manual." Accessed March 31, 2012. <http://download.intel.com/products/processor/manual/325462.pdf>.

Leat, Chris. "Forensic analysis of Mac OS X memory." Master's thesis, University of Westminster, 2011.

- Lee, Kyeongsik. "volafox: Memory analyzer for Mac OS X & BSD." Accessed May 24, 2012.
<https://code.google.com/p/volafox/>.
- Ligh, Michael, Steven Adair, Blake Hartstein, and Matthew Richard. *Malware analyst's cookbook and DVD: Tools and techniques for fighting malicious code*. Indianapolis: Wiley, 2011.
- Makinen, Juho. "Automated OS X Macintosh password retrieval via FireWire." Last modified February 29, 2008. <http://www.juhonkoti.net/2008/02/29/automated-os-x-macintosh-password-retrieval-via-firewire>.
- Malard, Arnaud. "H@CKMacOSX: Tips and tricks for Mac OS X hack." Last modified August 3, 2011.
<http://sud0man.blogspot.com/2011/08/hackmacosx-tips-and-tricks-for-mac-os-x.html>.

Institutional and Inter-Organizational Initiatives in Digitization

Digitization of Documentary Heritage Collections in Indic Language

Comparative Study of Five Major Digital Library Initiatives in India

Anup Kumar Das

*Centre for Studies in Science Policy, School of Social Sciences, Jawaharlal Nehru University
anupdas2072@gmail.com, anup_csp@mail.jnu.ac.in*

Abstract

Documentary heritage collections in Indic languages have been soul of indigenous digital libraries in South Asia. Some of the digital preservation initiatives in India received global acceptance are namely, Digital Library of India, Panjab Digital Library, Kalasampada Digital Library—Resource for Indian Cultural Heritage, National Databank on Indian Art and Culture, Traditional Knowledge Digital Library, and National Mission for Manuscripts, due to uniqueness in their collections and approaches. These projects also help in preserving socio-linguistically diverse cultural contents and achieving a sense of unity while online accessing using common platforms. This paper evaluates enrichment of collections and effectiveness of online platforms of five major digital library initiatives in India.

Author

Dr. Anup Kumar Das is an advance researcher and library professional attached with the Centre for Studies in Science Policy at Jawaharlal Nehru University (JNU), India. He was a UNESCO Consultant in the areas of Information for All and Open Access to Knowledge at its New Delhi office. He was awarded Ph.D. degree from the Jadavpur University, India in 2008 for his doctoral thesis titled “*An Evaluative Study of Some Selected Libraries in India undergoing the Process of Digitization.*” He has published a number of papers in different journals and presented papers in different international/ national conferences. His recent work includes ‘*Open Access to Knowledge and Information: Scholarly Literature and Digital Library Initiatives—the South Asian Scenario*’, UNESCO, New Delhi, 2008. A list of his publications and presentations is available at <http://anupkumardas.blogspot.in>.

1. Introduction

India is the country of ‘unity in diversity’, where multicultural society embraces a diverse people with anthropologically different linguistic, religious, ethnic, demographic and regional backgrounds. Here cultural diversity and cultural pluralism are coexisting for the centuries. India is the country of origin of many legendary ancient literature that form rich collections of Asian documentary heritage. South Asia also plays a significant role in shaping up world literary traditions.

Over the time Indian cultural institutions became the repositories of rich collections of cultural heritage resources embracing culturally and linguistically diverse communities across states of India. While traditional knowledge of linguistically diverse communities is largely un-documented, there were several attempts to collate them. Systematic documentation of traditional knowledge is centuries old practice of scholars and researchers to make knowledge re-usable by future the generations. These documentation initiatives ended up with producing literature of various kinds. On the other hand, some of the documentary heritage resources available with Indian institutions are on the verge of extinction due to lack of preservation and conservation initiative at the institutional level.

As a member country of UNESCO, India became de-facto signatory of the *UNESCO Universal Declaration on Cultural Diversity*, adopted unanimously by the UNESCO General Conference at its 31st

session held on 2 November 2001. This is an international standard-setting legal instrument which raises cultural diversity to the rank of “common heritage of humanity” [UNESCO, 2001]. The Declaration attempts to respond to two major concerns: (i) to ensure respect for cultural identities with the participation of all peoples in a democratic framework, and (ii) to contribute to the emergence of a favourable climate for the creativity of all, thereby making culture a factor of development.

The Article 6 of the *UNESCO Universal Declaration on Cultural Diversity* emphasizes on the equitable access to culturally diverse multilingual contents with help of digital technologies. Modern information and communication technologies (ICT), including internet technologies, have tremendous potentials to act as enabler for intercultural dialogue through digital dissemination of cultural information, particularly with culturally diverse contents. Cultural informatics can also bridge linguistically diverse contents through translations and adaptations. Thus, cultural informatics can help in making culture a factor of development.

UNESCO and its member states adopted an action plan for the implementation of the *UNESCO Universal Declaration on Cultural Diversity*. Some of main lines of the action plan embraced digital technologies for strengthening the access to diverse cultural resources available across the country. Text Box 1.2 provides a list of selected main lines of action plan related to access and dissemination, for implementation of the Declaration. As a member state of UNESCO, India is committed herself to take active role in the main lines of action plan for implementation of the Declaration [MCIT, 2004]. Several digitization and digital library projects in India, as documented in a recent UNESCO publication “Open Access to Knowledge and Information: Scholarly Literature and Digital Library Initiatives—the South Asian Scenario” and indicated in Table 1, have created an appropriate atmosphere for intercultural dialogue and intercultural partnership [Das, 2008]. These public-funded initiatives have tried to digitally include different communities in India, thus, further enhancing scope of cultural diversity and cultural pluralism in India [Ghosh, 2007].

The digital library initiatives in India have tried to contribute towards achieving multicultural and cross-cultural dialogs in a democratic society, in addition to making the endangered documentary resources digitally available. Digitization of documentary heritage collections in a culturally rich and diverse country is a major challenge to the ICT professionals and policymakers, due to nature of vastness versus available financial resources and institutional frameworks [UNESCO, 2009]. Thus, scaling up is real concern in India that needs to involve all possible stakeholders as well as end users.

Text Box 1. Article 6 of the UNESCO Universal Declaration on Cultural Diversity.

Towards Access for All to Cultural Diversity

“While ensuring the free flow of ideas by word and image care should be exercised that all cultures can express themselves and make themselves known. Freedom of expression, media pluralism, multilingualism, equal access to art and to scientific and technological knowledge, including in digital form, and the possibility for all cultures to have access to the means of expression and dissemination are the guarantees of cultural diversity.”

Text Box 2. Select Main Lines of Action Plan for the Implementation of the
UNESCO Universal Declaration on Cultural Diversity.

- Encouraging "digital literacy" and ensuring greater mastery of the new information and communication technologies, which should be seen both as educational discipline and as pedagogical tools capable of enhancing the effectiveness of educational services.
- Promoting linguistic diversity in cyberspace and encouraging universal access through the global network to all information in the public domain.
- Countering the digital divide, in close cooperation in relevant United Nations system organizations, by fostering access by the developing countries to the new technologies, by helping them to master information technologies and by facilitating the digital dissemination of endogenous cultural products and access by those countries to the educational, cultural and scientific digital resources available worldwide.
- Ensuring protection of copyright and related rights in the interest of the development of contemporary creativity and fair remuneration for creative work, while at the same time upholding a public right of access to culture, in accordance with Article 27 of the Universal Declaration of Human Rights.

2. Digital Library Development in India

Several scholars evaluated growth and evolution of digital library initiatives in South Asia [Das, 2008; Ghosh, 2007; Mittal, 2008; Mahesh, 2008]. Many of them feel that availability of scholarly e-resources supplements the universal access to knowledge. While commercial publishers had started offering e-

Development of Indian digital libraries and online repositories were planned when internet got penetrated into renowned academic and research institutions in 1990s. Indian institutions, such as Indian Institute of Science Bangalore also planned their open access digital repositories in late 1990s [Ghosh, 2007]. At the same time, Universal Digital Library was planned at global level to make available a million books online in free access mode. Its national surrogate Digital Library of India was also planned during the same time. Other digital libraries were also opened up to expand access to culturally-rich and linguistically-diverse contents to promote inter-cultural dialogues and to preserve literary assets of local and learned communities. In the last decade, many state and non-state agencies have taken significant initiatives in digitization and preservation of documentary heritage collections available across the country and also with neighboring South Asian countries.

Table 1 shows an indicative list of multilingual digital library initiatives in India. Some of them are early starters in offering online access to Indian documentary heritage collections to global communities. Protecting and safeguarding documentary heritage of local communities and long-term preservation of endangered documentary collections are main objectives of these digital library projects.

These digital libraries could also sensitize policymakers in taking affirmative actions in content localization. Thus, many Indic language literatures got digitized and archived in searchable of digital libraries in India. Indian Ministry of Communications and Information Technology (MCIT) has established a National *Digital Libraries* Cell within the Department of Electronics and Information Technology (DeitY) to streamline development of digital libraries in India. Indian Ministry of Culture has also been involved in supporting digitization of cultural contents including manuscripts and archival

Table 1. Indicative List of Multilingual Digital Library Initiatives in India.

Name of the Initiative	Implementing Agency	Funding Agency	Website
Digital Library of India (DLI)	Indian Institute of Science; IIIT Hyderabad; C-DAC	MCIT and others	http://www.new1.dli.ernet.in http://www.new.dli.ernet.in http://dli.cdacnoida.in
Kalasampada: Digital Library Resources for Indian Cultural Heritage (DL-RICH)	IGNCA	MCIT	http://www.ignca.nic.in/dlrich.html
National Databank on Indian Art and Culture (NDBIAC)	IGNCA	MCIT	http://ignca.nic.in/ndb_0001.htm
Kritisampada: National Database of Manuscripts	National Mission for Manuscripts, IGNCA	Ministry of Culture	http://www.namami.org/pdatabase.aspx
Panjab Digital Library (PDL)	Panjab Digital Library	Nanakshahi Trust and others	http://www.panjabdigilib.org
Digital Repository of WBPLN (DR-WBLLN)	West Bengal Public Library Network (WBPLN), CDAC Kolkata	Directorate of Library Services, West Bengal	http://dspace.wbpublibnet.gov.in/dspace/
Archives of Indian Labour (AIL)	V. V. Giri National Labour Institute & Association of Indian Labour Historians	Ministry of Labour	http://www.indialabourarchives.org/
Muktabodha Digital Library	Muktabodha Indological Research Institute	Donations from Individuals & Trusts	http://muktalib5.org/digital_library.htm
Traditional Knowledge Digital Library (TKDL)	Council of Scientific and Industrial Research (CSIR)	Department of Ayurveda, Yoga & Naturopathy, Unani, Siddha and Homoeopathy (AYUSH)	http://www.tkdli.res.in/
National Science Digital Library	NISCAIR, India	Council of Scientific and Industrial Research (CSIR)	http://nsdl.niscair.res.in/
Vigyan Prasar Digital Library	Vigyan Prasar, India	Department of Science and Technology	http://www.vigyanprasar.gov.in/digilib/

materials. Majority of digital library initiatives in India, as indicated in Table 1, are public funded and targeted at common citizens.

Table 1 highlights two multimedia digital library initiatives of the Indira Gandhi National Centre for the Arts (IGNCA), funded by MCIT, namely Kalasampada: Digital Library Resources for Indian Cultural Heritage (DL-RICH) and National Databank on Indian Art and Culture (NDBIAC).

Archives of Indian Labour (AIL) is a multimedia digital library, initiated by V. V. Giri National Labour Institute and Association of Indian Labour Historians. It has very unique collection of documents of Indian labour movements that include reports, photographs, and oral testimonials.

Panjab Digital Library (PDL) is a global initiative of Nanakshahi Trust and other institutions to archive culturally rich collections of Punjabi literature and historic documents on Sikhism.

Digital Repository of West Bengal Public Library Network (DR-WBPLN) is an important multilingual digital library, established using popular content management software (CMS) DSpace. It provides access to more than 10,000 digitized rare books and government reports. Majority of them are written in Bengali language.

Muktabodha Digital Library is an initiative of Muktabodha Indological Research Institute to preserve rare Sanskrit manuscripts and texts in multiple digital formats, and make them worldwide accessible through its website for study purpose. Currently it has texts from the Kashmir Shaivism, Tantra Shastra, Kaula-Trika, Saiva-Siddhanta, Virashaiva, Pancaratra, Shree Vidya, Shakta and Natha Yoga schools.

Traditional Knowledge Digital Library (TKDL) is a flagship initiative of Council of Scientific and Industrial Research (CSIR). It attempts to curb bio-piracy and misappropriation of formulations of Indian systems of medicine (Ayurveda, Yoga, Naturopathy, Unani and Siddha). Access to TKDL is given to patent offices around the world, for sensitizing them on the prior art in formulations of Indian systems of medicine.

Table 1 also shows names of some educational digital libraries, such as National Science Digital Library (NSDL) and Vigyan Prasar Digital Library (VPDL). NSDL is an initiative of NISCAIR, New Delhi (National Institute of Science Communication and Information Resources) and VPDL is an initiative of Vigyan Prasar. Both institutions publish books on popular science and science education. NSDL and VPDL provide access to digitized e-books of popular science, published by NISCAIR and Vigyan Prasar respectively. NSDL additionally generated contents on science education topics. VPDL provides access to digitized e-books in English as well as in Hindi language.

For the convenience of this study, five important digital library initiatives are briefly evaluated in the following sections. Selected initiatives are DLI, DL-RICH, PDL, AIL and DR-WBPLN. They represent linguistically and culturally diverse documentary heritage collections. All of them maintain open access web portals to facilitate web-based visitors make use of these unique collections. This study also identifies technical challenges common users face during their quest for knowledge discovery using these platforms.

2.1 Digital Library of India Project

Digital libraries development in Indian sub-continent got accelerated with the initiation of Universal Digital Library (UDL) or Million Books Project (<http://www.ulib.org/>). The Universal Digital Library is a global collaborative project, initiated in 2001 by the Carnegie Mellon University and its partner institutions in China and India. Initially, UDL got support from the US National Science Foundation.

Later, each participating country secured funding from their respective national government. In India, Indian Institute of Science, Bangalore in collaboration with Centre for Development of Advanced Computing (C-DAC), and International Institute of Information Technology (IIIT), Hyderabad initiated Digital Library of India (DLI) project in 2002 as spin-off of Universal Digital Library project. The Ministry of Communications and Information Technology (MCIT) of Government of India extended full financial support for this initiative. DLI project started establishing Regional Mega Scanning Centres (RMSCs) and other scanning centres across the country for scaling up digitization of rare books, rare periodicals and other literature, including manuscripts and copyright-free or out-of-print books. DLI

established five RMSCs across the country at Hyderabad (IIIT Hyderabad), Kolkata (C-DAC Kolkata), Allahabad (IIIT Allahabad), NOIDA (C-DAC Noida) and Bangalore (IISc Bangalore).

All RMSCs started networking with source libraries in their localities for obtaining the books required for digitization. Majority of books selected for digitization were Indic language books published in all official languages of India, including in English. As on 31st August 2012, contents available in top six languages are respectively English, Sanskrit, Hindi, Telugu, Bengali and Urdu covering about 91.3% of books available with DLI web portal at www.new1.dli.ernet.in.

DLI maintains two web portals, one at IISc Bangalore (www.new1.dli.ernet.in and its mirror www.new.dli.ernet.in) and one at C-DAC Noida (<http://dli.cdacnoida.in>). Earlier, IIIT Hyderabad maintained another DLI web portal (<http://dli.iiit.ac.in>). Later this portal merged with the DLI site (www.new1.dli.ernet.in).

Scholars associated with DLI project, indicated that DLI developed a significant amount of digitized Indic language contents, covering all major Indian language. Thus, DLI becomes a testbed for Indian language technology, facilitating development of OCR (optical character recognition), TTS (text-to-speech) and other related software for Indian language computing [Balakrishnan, 2006].

In a digitization project, the scanned pages are stored in image format. Images of textual pages can be converted into computer readable, full-text searchable and editable textual documents, if any OCR software is readily available for the respective language. Rate of accuracy in commercially available OCR software is usually above 90% for standard printed texts [Wikipedia, 2012]. In DLI portals, Indian language full-text contents are available in TIFF image format only, not in textual format. This is because no OCR software was used for converting Indic language texts into textual format, due to non-availability of efficient OCR software for Indian contents. But, documents in English and other European languages were converted into textual format. Pages of English language books are stored and retrieved in image as well as textual formats. However, research team of DLI project has tried development of OCR technologies for Indic languages on experimental basis using corpora of scanned images of Indian language documents. C-DAC, one of DLI project partners, has recently launched Chitrakan OCR software for Devanagari and Bengali scripts, covering Hindi, Sanskrit, Marathi, Nepali, Bengali, Manipuri and Assamese languages. Chitrakan is the first OCR system for Indian Languages. This software is now expanding to other Indian languages as well.

2.1.1 Retrieving Documents from DLI portals

DLI portals provide browsing as well as searching facility from its homepage. Online database search can be performed using author, range of years, subject, language, and name of scanning centre. A user can use a single search term or a combination of search terms, such as author and language. DLI does not provide metadata in an Indian language. Metadata for Indian language book is entered in transliterated form. For, example इतिहास is rendered as Itihas in the database. User has to enter words in transliterated form in the search box to perform the search. Then, a search result is displayed with titles of books and other information. When user selects a title of the book, metadata of that book is displayed onscreen. Then he can read the book online, page by page, as tiff, or rtf, or txt or html file. As Indian language books are not OCR'd, only tiff image can be viewed onscreen. For English language books, tiff, or rtf, or txt or html file for each page will be displayed and can be browsed page by page.

Similarly, a user can browse books from homepage of DLI portal by selecting title beginning with a letter (A to Z), or author's last name (A to Z), or range of years, or a subject, or a language or a source

library. Finally, user will reach the book of his choice and metadata of that book is displayed onscreen. Then he can read the book online, page by page, as tiff, or rtf, or txt or html file.

DLI portal requires a plugin to view a document page-by-page that makes reading books online a difficult proposition to occasional users of DLI portal. Modern, tech-savvy young users of DLI sometimes feel frustrated and critical to its presentation of page-wise view. It is contrary to present standard of reading e-books or chapters of an e-book online in PDF format without time-consuming process of downloading each and every page of a book. Users also feel temptation to download books and read at leisure, even using a tablet computer or a laptop. A user has recently developed a downloader software named “@ABS DLI Downloader” (present version 2.2) for downloading all tiff pages of the book from DLI portal in a batch to user’s computer for reading in a sequential manner [Shukla, 2009]. This saves the time of end user and helps him read the book at his convenience, without compromising with time for re-downloading.

A user feels that DLI provides shabby interface for reading books online. Its required plugins, namely Alternatiff/ Plugger plugin (for BookReader-1) and QuickTime plugin (for BookReader-2), if not available with them, also impede many users to read books at leisure [Nadig, 2007].

2.2 Kalasampada: Digital Library Resources for Indian Cultural Heritage (DL-RICH)

DL-RICH is an attempt to digitize contents of cultural resources available with IGNCA and its partner institutions across the country. IGNCA’s mandate is to prepare national inventory and documentation of different areas of cultural studies, such as, intangible cultural heritage, tangible cultural heritage, performing arts, folk arts, manuscript traditions, handicrafts and Indology. It has three regional centres across India at Bangaluru for southern region, Guwahati for north-eastern region and Varanasi for northern region. IGNCA and its regional centres partnered with local scholars engage in documenting endangered cultural resources to unearth cultural resources hidden with local communities. Thus, IGNCA has developed a vast collection of cultural resources, including rare books, manuscripts, old photographs and handicrafts.

DL-RICH was a first phase of digitization project of IGNCA to cover its existing collections as well as collections of its partner institutions. Over time, DL-RICH became a multimedia and multilingual digital library covering texts, images, audio and video recordings, 3D artefacts. DL-RICH received funding support from the MCIT. Majority of DL-RICH collection is accessible in offline mode within the premises of IGNCA and its regional centres. Only a part of its digitized collection is made freely accessible through DL-RICH web portal. This portal provides access to different segments of its collection with English interface and English transliterated metadata information. Further, a part the DL-RICH collection also got translated into Hindi and metadata information got transliterated into Hindi to provide access to selected collection online through the interface of CoIL-Net (Content Development in Indian Language Network), a project sponsored by MCIT. This CoIL-Net collection recently went offline temporarily, probably for maintenance work.

2.2.1 National Databank on Indian Art and Culture (NDBIAC)

IGNCA started second phase of digitization project as National Databank on Indian Art and Culture (NDBIAC), supported by MCIT and Archeological Survey of India (ASI). NDBIAC provides access to digitized images and audio-visuals provided by ASI and state archaeology departments. A user can access information on archaeological sites in different states of India. An archaeological site information usually contains a brief description, images and video clips. NDBIAC will provide information on all major

archaeological sites in India. NDBIAC also gives access to virtual walkthroughs of archaeological monuments, presently seven. The number will be increased in coming years.

In this phase of project IGNCA has started digitizing ASI publications and rare books from ASI Library. Some of the important digitized works include all back issues of ASI journal “Indian Archaeology - A Review,” ASI reports, rare books in Indic languages (Hindi and Sanskrit) and English. NDBIAC provides free access to many digitized books and reports from its web portal, <http://www.ignca.nic.in/asp/searchBooks.asp>.

2.2.2 National Database of Manuscripts ‘Kritisampada’

IGNCA is also hosting National Mission for Manuscripts (NMM), a national apex body for preservation and conservation of manuscript resources in the country. NMM maintains a National Database of Manuscripts named ‘Kritisampada’, an outcome of its nation-wide inventory survey of manuscripts. NMM is also working towards development of a national digital library of manuscripts. NMM has identified 45 collections of Manuscript Treasures of India (MTI), or *Vijñānanidhi*. These are very unique and rare collections of manuscripts. MTI will be given first preference in the next phase of digitization and will be digitally archived through NMM’s national digital library of manuscripts. List of 45 MTI collections is available at [www.namami.org/manuscript Treasures.htm](http://www.namami.org/manuscript%20Treasures.htm).

NMM also maintains a network of manuscript repositories for widening bibliographic control and safeguarding manuscript collections available with local communities and institutions across the country. NMM partners are classified as Manuscript Resource Centres (MRCs), Manuscript Conservation Centres (MCCs), Manuscript Partner Centres (MPCs) and Manuscript Conservation Partner Centres (MCPCs). NMM established a network of 47 MRCs, 32 MCCs, 32 MPCs and more than 200 MCPCs across the country for identifying, inventorying, preservation and conservation of endangered documentary heritage collections available in the form of manuscripts.

Indian government has played active role in bringing in international recognition of Indian documentary heritage collections and more specifically endangered manuscripts collections. Almost every year, India government nominates endangered and important documentary heritage collections for inscription on the Memory of the World Registers. A list of India’s nominated collections that got selected for inscription is indicated in Table 2. After inscription, the nominating institutions have started digital preservation of their commendable documentary collections. Most of these digitally preserved collections are kept in computer databases in local servers within the institutions, or stored on CD-ROMs, and DVDs. These are kept for onsite consultation by the privileged scholars who can afford to visit those institutions. Only two of them are partially available in a public domain digital library, namely, Saiva Manuscript in Pondicherry and I.A.S. Tamil Medical Manuscript Collection, which are partially available on the Muktabodha Digital Library website (http://muktabodha.org/digital_library.htm). All institutions have preferred the medium of microfilming for long-term preservation of these collections. However, their availability remains limited to institutions’ time and space. Publishing them as printed books is also an option that was exercised by a few institutions.

Interestingly, NMM has partnership with all Indian institutions mentioned in Table 2, with playing role of a Manuscript Resource Centre (MRC), Manuscript Conservation Centre (MCC), Manuscript Partner Centre (MPC), or Manuscript Conservation Partner Centre (MCPC). All inscribed items of MoW Register from India, except the Archives of the Dutch East India Company, have been recognized by NMM as Manuscript Treasures of India (MTI).

Table 2. Inscription of India's Documentary Heritage Items on MoW Register.

Name of Item	Year of inscription	Host Institution	Role with NMM	Whether Further Action Taken
The I.A.S. Tamil Medical Manuscript Collection	1997	Institute of Asian Studies, Chennai	MTI, MPC	Microfilmed & Digitized; Published books
Archives of the Dutch East India Company	2003	National Archive of the Netherlands	Nil	Microfilmed & Digitized; Published books
Saiva Manuscript in Pondicherry	2005	French Research Institutions in Pondicherry	MTI, MRC	Microfilmed & Digitized; Parampara CD-ROM, Digital Library; paper transcripts
Rigveda	2007	Bhandarkar Oriental Research Institute, Pune	MTI, MRC	Microfilmed & Digitized for onsite consultation
Tarikh-E-Khandan-E-Timuriyah	2011	Khuda Bakhsh Oriental Public Library, Patna	MTI, MRC, MCC	Microfilmed & Digitized for onsite consultation
Laghukālacakratantrarājatikā (Vimalaprabhā) [Buddhist Tantric literature]	2011	Asiatic Society, Kolkata	MTI, MPCC	Microfilmed & Digitized for onsite consultation

2.3 Multilingual Digital Libraries having Indic Language Digital Collections

Many institutions have attempted designing and development of digital archives profiling their unique collections of documentary resources. These institutions possess a significant number of rare and out-of-print books, periodicals and old photographic collection. They identify valuable source materials from their in-house collections for digitization. Sometimes they network with other libraries and document repositories to select source materials for digitization.

Table 3 shows an indicative list of digital libraries and digital repositories having collection of Indic language documents. This Table also indicates that some digital libraries use content management software (CMS), such as DSpace and Greenstone. DSpace and Greenstone are popular free and open source software (FOSS) for building digital libraries. CMS helps in developing a structured digital library with standard sets of metadata and facility of cross-searching by external metadata harvesters. Many Indian institutions have developed institutional repositories using DSpace and EPrints software.

Digital Repository of West Bengal Public Library Network (DR-WBPLN) is a joint effort of public libraries of West Bengal under the Directorate of Library Services and C-DAC Kolkata. This repository was created at the 'Heritage Preservation Unit' of West Bengal State Central Library in technical collaboration with C-DAC Kolkata. C-DAC has digitized significant number of rare books, available with State Central Library and its associated libraries. This portal is developed using DSpace software. Full-text contents of this repository are available in PDF format. This repository also provides access to digitized government publications and gazettes. A partial set of metadata information is available both in Bengali language and English, e.g., Title of book. Availability of metadata information set in Indic language is indicated in Table 3. Other metadata information is available in English only, e.g., name of

Table 3. Indicative List of Indian Digital Libraries having Indic Language Contents.

Name of Digital Library	Organization	Indic Language Collections	CMS used	Whether Metadata in Indic Language Available	Metadata in Indic Language
Digital Repository of W.B. Public Library Network http://dspace.wbpublibnet.gov.in:8080/dspace	West Bengal State Central Library & CDAC Kolkata	Digitized Rare Books (10195 items, about 80% in Bengali)	DSpace	Yes, Partial	Title, Appears in Collections, Description
Panjab Digital Library http://www.panjabdigilib.org/	Panjab Digital Library; Nanakshahi	Manuscripts 704; Digitized Books 994; Magazines 432; Newspapers 540; Photographs 103	-	No	-
Archives of Indian Labour (AIL) http://www.indialabourarchives.org/	V. V. Giri National Labour Institute & Association of Indian Labour Historians	Digitized Reports, A-V materials, Images.	Greenstone	No	-
Digital Repository of VPM http://dspace.vpmthane.org:8080/jspui/index.jsp	Vidya Prasarak Mandal, Thane	Marathi E-Books; Marathi E-Journals (373 items).	DSpace	Yes, Partial	Title, Authors, Publisher, Appears in Collections
E-Gyankosh http://www.egyankosh.ac.in/	Indira Gandhi National Open University, New Delhi	Digital learning resources (25025, about 1% in Hindi and other Indic lang.).	DSpace	No	Title, Appears in Collections
ASI Digital Library http://www.ignca.nic.in/asp/searchBooks.asp	ASI Library; IGNCA New Delhi	Digitized Rare Books, ASI Publications (out-of-print), journal <i>Indian Archaeology</i> .	-	No	-

Data as on 1 September 2012

author, name of publisher. This repository helps in outreaching Bengali literature to global communities of scholars and enthusiastic book readers.

Panjab Digital Library (PDL) is a significant digital library initiative jointly created by Punjabi communities of India and abroad. The mission of PDL is to locate, digitize, preserve, collect and make accessible the accumulated wisdom of the Panjab region. PDL has been digitizing significant number of rare books, magazines, newspapers, photographs and manuscripts available with Punjabi communities in the state of Punjab and Sikh religious institutions (Gurdwaras). PDL is freely accessible to the registered users from its web portal at www.panjabdigilib.org. The documents can be browsed or searched after selecting its category, namely, manuscripts, books, magazines, newspapers and photographs. Full-text

contents of this digital library are available in image format and can be viewed page-wise. Table 3 also indicates that metadata information of PDL documents is available in English in transliterated form. This digital library helps in outreaching Gurmukhi and Punjabi literature to global communities of scholars and vivid book readers. PDL project was bestowed the Manthan Awards in 2010 in the category of E-Culture & Heritage.

Archives of Indian Labour (AIL) is another significant initiative for building multilingual and multimedia open access repository. AIL was set up in the month of July in 1998 as a collaborative project of V.V. Giri National Labour Institute and the Association of Indian Labour Historians. This portal is developed using Greenstone software. Presently, it provides access to 12 unique collections related to labour movements in different industries, reports of labour commissions, publications of labour unions, transcripts of recordings of oral history. AIL is freely accessible to its users from its web portal at <http://www.indialabourarchives.org/>. The documents can be browsed or searched after selecting a collection. Documents are either textual or images. Digitized full-text contents in English language of this digital repository are available in HTML format, where as digitized documents in Indic language are stored in image format and can be viewed page-wise. Table 3 also indicates that metadata information of AIL documents is available in English in transliterated form. This digital archive helps in outreaching history of South Asian labour movements and social life of industrial workers to global communities of historians and social scientists. AIL project was bestowed as the pioneering and predecessor of all recent digital library initiatives of Indian sub-continent.

Table 3 also indicates other digital repositories providing access to digitized Indian contents. Digital Repository of Vidya Prasarak Mandal provides full-text access to some Marathi books and magazines. E-Gyankosh of Indira Gandhi National Open University (IGNOU) provides full-text access to digitized course materials in Indian language, although majority of IGNOU course materials are available in English. Similarly, ASI Digital Library of Archeological Survey of India provides full-text access to selected digitized rare books of ASI Library collections. As indicated earlier ASI Digital Library is a part of NDBIAC project of IGNCA.

Table 4 indicates another set of multilingual digital libraries, established by institutions abroad in collaboration with South Asian partner institutions. These digital libraries have significant collections of digitized contents published from South Asia. These also have contents in Indic language. Majority of contents of these digital libraries is focused on South Asian studies, South Asian literature and Indology.

Table 4. Overseas Digital Libraries having Sourcing Partners
in South Asia/ Collections from South Asia.

Name of the Initiative	Implementing Agency	Funding Agency	URL
Digital Himalaya	Digital Himalaya Project team	University of Cambridge	http://www.digitalhimalaya.com/
Tibetan and Himalayan Library (THL)	University of Virginia Library; Institute for Advanced Technology in the Humanities, USA	University of Virginia	http://www.thlib.org/
The Digital South Asia Library	University of Chicago and the Center for Research Libraries, USA	University of Chicago	http://dsal.uchicago.edu/

3. Conclusion

Indian multilingual digital library initiatives have shown keen interests in ‘lean backward’ to digitize important documentary heritage collections available with Indian institutions. Most of these items have status of rare, out-of-print or copyright-free books and documents. After completing the process of digitization, these rare books are usually archived in an online digital library platform. Educated common people consult the open access digital libraries to read digitized books online and even download some of them for future reading. A rare book in an Indic language gets much attention to the general book readers, because readers get easy access to a full document. Popular literature if available with a digital library platform, it will increase the chance of web visibility of respective digital library. On the other hand, documentary heritage collections in the form of manuscripts are of special interests of scholars or subject specialists.

All five digital libraries evaluated in this study have sufficient propositions to become popular choices. Digital Library of India has largest digital collections and its technical glitches are also higher than other initiatives due to flaw in design. DLI also has many dead web-links showing pages do not exist while browsing, which needs to be corrected.

Present study tries to focus availability of contents of documentary heritage collections in online platforms. After evaluating national and institutional initiatives, we still believe that majority of digital library initiatives failed to archive all their digitized contents at fullest extend. While international and national policy instruments are in place to augment access to cultural diversity of a large country like India, most of the initiatives are short-focused, in terms of incorporating interoperable, cross-searching and metadata harvesting functionalities. Many of the online resources could not get global visibility due to low level of outreach, advocacy and awareness raising activities.

This paper indicates some digital libraries provide metadata information in Indian language or in transliterated English, for the documents in a respective Indian language. Lack of searchable metadata in Indian language hampers readers’ quest for knowledge discovery in a digital environment. It also makes documents inaccessible to some prospective readers. Gradually, we need to focus on searchable metadata information in Indian language.

Multilingual digital library initiatives in India have helped in bridging digital divide in the country by making Indian language documents freely available to the masses and pushing content localization efforts of associated online platforms.

India has gained momentum in open access movement, by establishing open access channels of digital publishing. Many authors writing in vernacular languages in India would be interested in putting documents in Indian digital libraries with creative commons licenses. Thus, we now need to focus on “lean forward” to include born digital contents in multilingual digital library collections.

References

- Balakrishnan, N., Raj Reddy, Raj Madhavi Ganapathiraju, and Vamshi Ambati. “Digital Library of India: A Testbed for Indian Language Research.” *TCDL Bulletin* 3, no. 1 (2006): 1-16.
- Das, Anup Kumar. *Open Access to Knowledge and Information: Scholarly Literature and Digital Library Initiatives – the South Asian Scenario*. New Delhi: UNESCO, 2008.
- Das, Anup Kumar, Chaitali Dutta, and B. K. Sen. “Information retrieval features in Indian digital libraries: a critical appraisal.” *OCLC Systems & Services* 23, no. 1 (2007): 92-104.

- Ghosh, S. B., and Anup Kumar Das. "Open Access and Institutional Repositories – A Developing Country Perspective: a case study of India." *IFLA Journal* 33, no. 3 (2007): 229-250.
- India, Ministry of Communications and Information Technology. *Digitization of Culture – Background Note for Asia IT Ministers' 2nd Summit, Hyderabad, 2004*.
- Mittal, Rekha, and G. Mahesh. "Digital libraries and repositories in India: an evaluative study." *Program: Electronic Library and Information Systems* 42, no. 3 (2008): 286-302.
- Mittal, Rekha, and G. Mahesh. "Digital Libraries in India: A Review." *Libri* 58 (2008): 15-24.
- Nadig, Hari Prasad. "Digital Library of India: Download all that you can...." 2007. Accessed 1 September 2012. <http://hpnadig.net/blog/index.php/archives/2007/02/22/download-all-that-you-can>.
- Shukla, Alok Bhushan. "Where Knowledge is Free – Digital Library of India." 2009. Accessed 1 September 2012. <http://alokshukla.wordpress.com/2009/12/11/where-knowledge-is-free-digital-library-of-india/>.
- UNESCO. "Universal Declaration on Cultural Diversity." 2001. Accessed 1 February 2009. <http://www.un-documents.net/udcd.htm>.
- UNESCO. *UNESCO World Report: Investing in Cultural Diversity and Intercultural Dialogue*. Paris: UNESCO Publishing, 2009.
- Wikipedia. "Optical character recognition." 2012. Accessed, 1 September 2012. http://en.wikipedia.org/wiki/Optical_character_recognition.

Digital Heritage Preservation - Economic Realities and Options

Ronald Walker

Executive Director, Canadiana.org

Abstract

The demand for digital heritage preservation is increasing, particularly in response to the demand for online access by professional and amateur researchers, family historians, Universities, K-12 educators. The traditional government grant as a source of funding for cultural heritage projects in general and digitization projects in particular, however, is increasingly rare in the current economic climate. These economic realities inspire new funding models for heritage collection digitization, perpetual preservation and online access. The focus of this paper is to discuss various strategies which may be available to not-for-profit institutions to achieve a sustainable financial basis for their operations. The paper explores the following strategies: Budgeting strategies; Grants and sponsors; Subscription models; Value added strategies; Content repurposing; Curated portals; Advertising; Centre of Excellence services; Sustainability Foundation.

Author

As Executive Director, Canadiana.org, Ron Walker has the mandate to create a Canada-wide collaboration of memory institutions to deliver digitization, preservation and access to Canada's documentary heritage. Before joining Canadiana, as a senior manager in government, founder and CEO of private information technology products and services corporations, business consultant, project manager and technology architect he has delivered several successful pan-Canadian and international projects.

1. Introduction

The traditional grant-based funds for cultural heritage projects are becoming increasingly rare in the current economic climate. At the same time, the demand for access to digital heritage is increasing, particularly by the general public, professional and amateur researchers, family historians, universities, and educators in primary and secondary schools. Modern users have come to believe that if the information they want is not online, it does not exist.

These economic realities are inspiring new funding models for heritage collection digitization, perpetual preservation and online access. This paper discusses various strategies which may be available to not-for-profit institutions to achieve a sustainable financial basis for their operations. The paper briefly explores the economic realities, value proposition and economic options, including:

- Commercial approach;
- Sustainability foundations;
- Budgeting strategies;
- Grants and sponsors;
- Subscription models;
- Value added strategies;
- Content repurposing;
- Curated portals;
- Advertising; and
- Centre of Excellence services.

2. Economic Realities

Documentary heritage is often physically fragile and practically inaccessible to most citizens. The most cost effective and increasingly the only way to provide wide access to collections is on the internet.

With the widespread growth in using the internet to access information we have come to expect information to be online. If we cannot easily access information online then they may assume, for all practical purposes, it does not exist or is not accessible. The old model of going to the library to do research has been replaced by the web. At the same time, the modern web user expects content to be free, so the demand for “open” (free) access to digital heritage is increasing. Someone, however, has to pay for it.

Memory Institutions holding documentary heritage collections often provide their own, or collaborate with others, to build and maintain digital heritage repositories for digitized and born-digital content. These institutions are typically publicly funded, and have relied on grants and/or re-directed internal budgets to achieve this.

In times of economic recessions, governments reduce and eliminate grants for cultural policy areas, including heritage digitization projects. There is a growing urgency to preserve unique library collections for future generations as government cuts close libraries and collections are moved into long-term warehouses, or are scattered and lost.

Digitization costs that require personnel are increasing, but technology costs for preservation and access are decreasing.

Commercial and non-profit organizations have entered the field providing free or low-price digitization services with new business models to pay for and exploit the value of online documentary heritage:

- subsidization from other business income generators;
- large scale public funding initiatives; and
- for-profit business models.

Summary - Traditional funding sources for digitization are becoming more uncertain, while demand for open and online access is increasing. The expenditure and funding models for Not-for-Profit digitizers and access providers need to be realigned according to the shifting financial landscape.

3. The Sustainable Digital Heritage Preservation Challenge

A digital heritage preservation institution must be able to maintain a perpetual, self-sustained operation. For digital heritage to be perpetually preserved, its content must be perpetually accessible.

A digital repository and discovery infrastructure will be in a constant state of change, adapting to new technologies and media standards. It requires:

- Infrastructure, policies, procedures and practices compliant with a high level of standard, including third party review, i.e., Trusted Digital Repository;
- Mutual multiple redundant backup across a network of institutions supporting Trusted Digital Repositories;
- Ongoing conversion of file formats to adapt to changing standards;
- Ongoing hardware and software infrastructure refreshing as technologies evolve;
- Ongoing storage media refreshing; and
- Ongoing evolution of discovery and access tools.

Summary - Digital heritage preservation requires an ongoing viable institution that meets the standards of a trusted Digital Repository, and is not susceptible to the vagaries of the economy, reliant on grants, sponsorship or the kindness of strangers.

4. Value Proposition

Primarily documentary heritage is a priceless treasure that must be preserved and made accessible forever. Documentary heritage has real, ongoing, useful value to current and future generations. Depending on the user's point of view, the value may be cultural, economic, academic, educational, legal, entertainment, etc.

The market can be defined with three primary customer groups, arbitrarily named heritage, history community and collection holders.

The **heritage group** includes institutions such as research libraries, universities and other heritage conscious organizations/governments/granting authorities. In Canada, well over 100 universities use online collections. Federal and provincial government departments also use online collections to research and provide services to their own employees.

The **history community** includes individuals generally interested in history, genealogy, family history, community history associations, and those who are interested in the way things were done in the past (for example, cookbooks, periodicals, etc.). In Canada, roughly based on the number of visits to our online collections and the number of subscribers to the Canadian History magazine, we estimate there are 450,000 (fondly named) history buffs cutting across all ages and genders. Of genealogical societies and individuals interested in family history there are more than 1.7 million Canadians who visit genealogy websites monthly and spend around \$250 a year for access to online genealogy related records.

Collection holders are those who hold collections and need digitization, preservation and access services.

We are witnessing a growing demand from the public for access to digital heritage. For example, Canadiana.org's Early Canadiana Online collection was developed for academics. In the last two years, after making it more accessible by the public, fully two-thirds of the page views have been from the public and one-third by academic researchers and university students. The following table shows the public's high interest subject areas.

Table 1: High Interest Subjects

Top 100 of Early Canadiana Online Collection	percentage of views
Native Studies	31%
Government Publications	28%
Women's History	20%
French Canadian History	11%
English Canadian History	8%
Hudson Bay Company	2%

Another example is the 1911 Canadian Census. When Library and Archives Canada (LAC) put the 1911 Canadian Census online, it averaged 17 downloads per second for the first year. The family history TV program, “Who do you think you are” was so popular, the LAC hosted website saw 90,000 - 100,000 visits each time the program aired. When Canadiana.org launched the Canadiana Discovery Portal, the public showed their interest with over 10,000 hits a day to the website. Clearly, along with academic research, Canadians want access to their heritage and expect to find it online.

Summary - While free and open access is the ideal model, the reality is that digital preservation is not free. If documentary heritage is valuable to a wide market, then documentary heritage can be made to pay for its own preservation.

5. Strategic Options

These economic realities inspire new funding models for heritage collection digitization, perpetual preservation and online access. Various planning strategies are available to not-for-profit institutions to achieve a sustainable financial basis for their operations.

5.1 Profit / Not-for-profit - Thinking like an entrepreneur

A financially sustainable not-for-profit organization needs to think and operate like a highly competitive for-profit corporation. One key difference is a non-profit corporation does not pay out as dividends to shareholders. When this strategy works, a competitive, successful not-for-profit needs to have a plan in place for when revenues exceed expenses within the taxation year. One long-term option is a “sustainability foundation”.

5.2 Sustainability Foundations - Thinking really long-term

A financial foundation is a common model used by not-for-profits with charitable status. A foundation provides a long-term, or even permanent financial vehicle to maintain and grow capital, and provide dividends for operations. This is perhaps the only model that can meet the criteria of perpetual sustainability. It is a long-term approach and will no doubt take a long time to fill the coffers of the foundation sufficiently to meet all the operating needs of a heritage preservation institution. Theoretically, the sustainability foundation could eventually fund all the operations costs, but it is healthier to assume a growing business strategy.

6 Revenue Strategies

6.1 Budgeting Strategies

Since grants typically do not provide for ongoing operations, the grant recipient institution is left with a new operations and maintenance budget item when the digitization project ends. This ongoing operations and maintenance obligation is susceptible to budget cuts in the future (and therefore program cuts), so perpetual preservation is not assured.

It is important that non-profit organizations ensure that on-going operational funds not be tied to granting sources of declining reliability. If grant money is used for projects, then make sure the operations and maintenance budget can absorb the new content resulting from that project.

6.2 Grants and Sponsors

Despite the warnings about relying only on grants and sponsors, both are still a very good source of funds for digital heritage preservation. Over the long-term, when grants are available, use them to subsidize content collection projects, reserving your operating budgets for ongoing operations. While grants are usually tied to a project, public and private sponsors will be more open to long-term preservation. A corporate sponsor may wish to gain marketing value with their logo on your website. Private sponsor's may enjoy having their name associated with preservation of specific content, whether it is a book, a sub-collection or a complete library.

6.3 Subscription models

The cost to build and preserve a digital heritage collection is substantial and ongoing. If the collection, by being online, provides a cost savings to institutions who need access to the collections, then an institutional subscription model is a supportable revenue source.

Depending on the content, private individuals, particularly those with special interests such as genealogy, historical research, cultural exploration, etc., will pay a subscription if the content is of interest, is not available or easily accessible elsewhere online, and it is more cost effective than traveling to research the original source materials.

6.4 Value added strategies

Open access, however, is still a desirable objective. If some or all content in the digital library is freely accessible, there is still an opportunity to offer premium services, for which some people will be happy to pay a fee. In other words, the content is free but enhanced access can be offered for a fee. Users have the option to buy the enhanced services to save time searching for relevant content.

Digital libraries can also add value to their collections by providing special curated views of their material, organized by specific historical themes, such as prominent historical events, geographic areas, cultural topics, occupational categories, etc. These thematic, curated collections can be offered as subscriber services.

6.5 Content re-purposing

Digital assets can be re-purposed and sold to raise funds for ongoing support. The micro-purchase model with eCommerce infrastructure can be built by the digital library or out-sourced to a distribution channel. Examples of re-purposed content include:

- Download searchable PDF;
- Download eBook reader compatible versions;
- Print and bind books and periodicals;
- Print and mount images as poster, frame, etc.; and
- Download audio and video files.

Re-purposed content might also increase the organization's chances of obtaining funding from private sponsors, particularly if the content is of special interest to a sponsor.

6.6 Advertising

Ads on your digital library website can generate substantial revenue if your site receives a very high volume of visits. On the negative side, ads can also destroy the image of your web site. There is a middle ground that can generate constant revenue with "tasteful" ads. Pages can be designed with areas for advertiser logos for which they pay a fixed amount regardless of the traffic. Typically this will be ads for products and services that your main users would relate to. This is overlapping with sponsor recognition and can be sold that way, giving the advertiser and web page a more stylish look. Extra features, e.g., expanding information boxes, links to further information, demo's, etc., can also be added for higher rates.

6.7 Centre of Excellence Services

Memory institutions that have been in the business digitizing, preserving and making accessible documentary heritage have usually developed significant expertise in these areas. This expertise can be used to provide digital documentation services to third parties in private industry or government organizations, again for an appropriate cost recovery based fee.

Examples of the marketable services include:

- Digitization / scanning;
- Ingesting already scanned and born digital;
- Metadata enhancement for pagination, article linking, etc.;
- OCR generation to metadata for keyword searching;
- Searchable PDF generation;
- Other derivative formats generation, e.g., eBooks;
- Cataloguing;
- Loading and indexing into a secure repository;
- Hosting;
- Web discovery portal development;
- Web portal hosting;
- Collection hosting;
- TDR Preservation Services;
- Backup services for other repositories; and
- Consulting services on all aspects of workflow and technology planning and support for digitization, preservation and access.

7. Conclusion

Preservation of digital heritage requires evolving technologies, and high standards of operations with the requisite policies and procedures; but, critical to delivering the preservation service is perpetual financial sustainability.

Digital heritage has significant financial, cultural and scientific value, and its preservation is important to the current and future generations of academics, researchers, and the general public of all levels of education.

Government grants for digitization projects are an excellent funding strategy, but do not help the perpetual financial stability of a preservation repository.

Not-for-profit organizations can both operate efficiently and receive compensation for services provided, including harvesting heritage assets, preserving, making accessible, repurposing for copy-on-demand and curated collections, subscription services to closed collections and providing optional premium services for open collections.

Preservation institutions can, by efficient operations and exploiting revenue generating strategies, invest excess revenue in foundations as a long-term financial strategy towards perpetual sustainability.

8. About Canadiana.org

Canadiana.org is a membership alliance governed by an active volunteer Board of Directors made up of distinguished scholars and representatives of major memory institutions from across Canada. The organization is pan-Canadian in outlook and governance and aims to represent the interests of many stakeholder constituencies, including content creators, content holders, and users of cultural heritage and research resources.

The mission of Canadiana.org is to support enduring access to Canada's digital documentary heritage for Canadians and the world. Canada's libraries, universities, museums, archives, and government agencies possess numerous rich digital collections containing our nation's documentary heritage. These collections are continually expanding and include many different types of content, including books, journals, newspapers, government documents, photographs, maps, postcards, sheet music, audio and video broadcasts.

Canadiana.org works together with partners to strengthen our collective ability to present Canada's documentary heritage content online. The organization acts as a coordinator, facilitator and advocate for digitization initiatives, along with providing digitization, preservation, and access services and infrastructures.

The prime objective of Canadiana.org is to provide online access to Canadian documentary heritage for researchers, historians, students, new Canadians, and for the public. Currently, less than 13% of Canada's printed knowledge materials is available online. Canadiana.org aims to make the published record of Canadian experience and creativity available to all, and to establish the collaborative network essential to doing this systematically, and to maintain this for the long-term.

Positioning Libraries in the Digital Preservation Landscape

S. K. Reilly

LIBER- the European Association of Research Libraries

Abstract

This paper draws on LIBER's experience in several European best practice network projects related to the digitisation of cultural heritage (Europeana Libraries, Europeana Newspapers) and to digital preservation (APARSEN). Collaboration with a diverse set of practitioners, including both public and private stakeholders, means that libraries can stake their place in the common vision for digital preservation and ensure that the issues surrounding the preservation of digital cultural heritage are represented in this vision. In the long-term, this collaboration should ensure that digitised cultural content from libraries is made available in the most cost effective and sustainable way to ensure continued access into the future.

Author

Susan is the LIBER Project Officer and manages LIBER participation in several EU projects in the areas of Open Access, data exchange, digital preservation and digitisation. She holds an MSc in Information Management from the University of Sheffield, and has several years experience working across a range of libraries, including management of the Library service at the Irish Management Institute in Dublin. Her interests range from open data, digital heritage, collaboration and innovation, and research infrastructures.

1. Introduction

The Stichting LIBER Foundation (Stichting LIBER)¹ is the principal association of the major research libraries of Europe. Its current membership includes 425 national and research libraries from more than forty countries, mainly but not only, in Europe. Its overall aim is to support a functional network across national boundaries in order to ensure the preservation of European cultural heritage, to improve access to collections in European research libraries, and to provide more efficient information services in Europe. In the 2009-2012 LIBER strategy,² (1) Digitisation and Resource Discovery, and (2) Heritage Collections and Preservation were identified as 2 of the 5 key performance areas for the Association.

Under digitisation and resource discovery LIBER set out to actively contribute to the digitisation of research library collections through the sharing of knowledge and best practice. In the area of Heritage Collections and Preservation the focus was on the development of standards, addressing access and preservation issues through and in digitisation, and the curation and preservation of born digital and digitised material.

For both key performance areas collaboration and participation in European projects were identified as vehicles for achieving their goals.

Over the past 3 years LIBER has worked towards developing its network and actively seeking out and building on opportunities to engage in collaborative projects to help develop best practice in digitisation and increase the visibility of library collections from across Europe.

¹ See <http://www.libereurope.eu>

² Making the case for European research libraries LIBER Strategic Plan 2009-2012, LIBER, accessed August 27, 2012, <http://www.libereurope.eu/sites/default/files/d5/LIBER-Strategy-FINAL.pdf>.

2. Libraries & the Digital Agenda in Europe

As increased access to knowledge in the Digital Economy is now being viewed as key to economic development, libraries are under pressure to increase access to the knowledge within their collections. Some national libraries³ are even redefining their functions to more explicitly state their role in supporting education and research.

Institutions such as the European Commission have great influence over the development of the Information Landscape. Initiatives such as the Digital Agenda⁴ have brought European research libraries into the centre of political discussions on the empowerment of the European citizen. This has presented LIBER with the opportunity to position research libraries at the centre of the Digital Agenda, in particular via their engagement in the dialogue surrounding development of European research infrastructures.

The Commission has also set some clear targets for the digitisation of Europe's cultural heritage.⁵ By 2015 all public domain masterpieces must be made available via Europeana and, by 2025, all of Europe's cultural heritage. As the requirement for the legal deposit of born digital becomes a reality, the size of such national library holdings will increase dramatically. Although progress has been unsatisfactory thus far,⁶ the Orphan Works Directive, in conjunction with Arrow,⁷ should open up space for the mass digitisation of and increased accessibility of cultural heritage material.

The Digital Libraries Initiative⁸ sets out to make all of Europe's cultural resources and scientific records—books, journals, films, maps, photographs, music, etc.—accessible to all, as well as preserve them for future generations. In its Recommendation on the Digitisation and Online Accessibility and Preservation of Cultural Heritage Material⁹ acknowledges that:

1. There is a need to increase the volume and types (e.g., audiovisual) content available in the Europeana Portal;¹⁰
2. That no clear and comprehensive policies are in place on the preservation of digital content in several EU Member States and the absence of which poses a threat to the survival of digitised material and may also result in the loss of material produced in digital format;
3. That digitisation is costly and steps should be taken to reduce cost through collaboration for economies of scale and public-private partnership.

The recommendations to address these issues include:

³ Connecting Knowledge: Scottish National Library Strategy 2011-2012, National Library of Scotland, accessed August 27, 2012, <http://www.nls.uk/media/896838/2011-2014-strategy.pdf>.

⁴ Digital Agenda for Europe: Communication from the Commission (26/08/2010), accessed August 25, 2012, [http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52010DC0245R\(01\):EN:NOT](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52010DC0245R(01):EN:NOT).

⁵ Digital Agenda: encouraging digitisation of EU culture to help boost growth, European Commission Press Release October 28, 2011, accessed August 26, 2012, <http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1292>.

⁶ Ben White, "Guaranteeing Access to Knowledge: The Role of Libraries," *WIPO Magazine* (August 2012), accessed August 25, 2012, http://www.wipo.int/wipo_magazine/en/2012/04/article_0004.html.

⁷ See <http://www.arrow-net.eu/>.

⁸ See http://ec.europa.eu/information_society/activities/digital_libraries/index_en.htm.

⁹ Commission Recommendation of 27 October 2011 on the digitisation and online accessibility of cultural material and digital preservation, European Commission, Accessed August 27, 2012, http://ec.europa.eu/information_society/activities/digital_libraries/doc/recommendation/recom28nov_all_versions/en.pdf.

¹⁰ See www.europeana.eu.

1. The use of domain specific cross border aggregators to bring about economies of scale;
2. Ensure the use of common digitisation standards defined by Europeana in collaboration with the cultural institutions in order to achieve interoperability of the digitised material at European level, as well as the systematic use of permanent identifiers;
3. Increased targets for digitisation activity;
4. Make the necessary arrangements for the deposit of material created in digital format;
5. In order to guarantee its long-term preservation.

The implication of these recommendations for Europe's libraries are that (a) there is a need for a library domain aggregator, (b) that they must engage in the development of digitisation and metadata standards, and (c) find new ways of reducing the cost of and funding digitisation, and, if they are mandated to preserve born digital material, (d) ensure that they have the correct skills and infrastructure in place to do so.

A survey carried out by the NUMERIC project¹¹ found that only 1% of library collections have been digitised but that a further 30% of the collection was going to be digitised. This represents a huge investment in digitisation, but to what end?

The New Renaissance Report,¹² which examines the digitisation of European cultural heritage, sees this investment as beneficial in terms of:

- wider access to and democratisation of culture and knowledge
- the educational system—both schools and universities
- the development of new technologies and services for digitisation
- for digital preservation
- interacting in innovative ways with the cultural material

Applying this logic to the library domain indicates that for libraries to ensure that full benefits of their digitisation activities are realized they must ensure that they maximize the visibility of their collections, not just to the general public but to those in the education system and must make it possible for individuals to interact with the content in new and innovative ways.

3. Collaboration

The first milestone in LIBER's work towards seeking opportunities to engage in relevant collaborative projects was its involvement in Europeana Travel. Europeana Travel¹³ was a 2 year project, which started in May 2009, funded by the European Commission. The aim of the project was to digitise content based around the theme of travel and make it available in Europe's portal for digitised cultural heritage, Europeana. Although LIBER was not an official partner in the project, Europeana Travel was a significant project for the Association because it brought together LIBER members from both national and university libraries as partners to develop best practice in digitisation. It also raised issues surrounding sustainability as the national and the other research libraries within the project used different

¹¹ Nick Pool, "The Cost of Digitising Europe's Cultural Heritage," A Report for the Comité des Sages of the European Commission, Collections Trust, 2010, accessed July 24, 2012, <http://www.collectionslink.org.uk/discover/sustaining-digital/739-the-cost-of-digitising-europes-cultural-heritage>.

¹² Maurice Lévy, Elisabeth Niggemann, and Jacques de Decker. *The New Renaissance* (Brussels: European Commission, 2011), Accessed April 24, 2012, http://ec.europa.eu/information_society/activities/digital_libraries/doc/reflection_group/final_report_%20cds.pdf.

¹³ See <http://www.europeanatravel.eu/>.

aggregators to deliver their content to Europeana. The Consortium of European National Libraries (CENL) had already established aggregation service, The European Library,¹⁴ but this service was only open to members of CENL. Europeana Travel set up a LIBER aggregation service for research libraries, which was to be extended to other libraries after the life of the project. As the project began to come to a close it became apparent that the best solution for sustaining an aggregation service for both groups of libraries would be to open up The European Library service to research libraries.

The Europeana Travel project also conducted a survey of the state of digital preservation practice in LIBER libraries who were digitising their collections. The survey¹⁵ found that there was an under investment in skills for digital preservation and often a lack of written policies on the preservation of digitised material. It recommended (1) the sharing of best practice and (2) investment in digital preservation infrastructures.

4. Sustainability

Europeana Travel established the basis for collaboration between library networks. The Europeana Libraries¹⁶ project brings together several European library networks, incorporating both national and research libraries, in order to create a partnership which will provide value to both researchers and research libraries alike. LIBER, CENL, the Consortium of European Research Libraries, and the Europeana Foundation are all partners in the project.

It addresses the issue of sustainability by opening up the national library aggregation service, The European Library, to research libraries. It uses this service to aggregate a critical mass of valuable content from European research libraries. By the end of the project in December 2012 over 5.1 million objects, including 1,200 film and video clips, 850,000 images and 4.3 million texts (books, journal articles, theses, letters) will have been ingested from 19 research libraries. Much of this content is full text and of particular value to researchers. To maximise on the potential of this content, the project also set out to develop full text search capabilities and a search portal that provides tools specific to research.

The projects value not only lies in the creation of a single aggregation service for libraries, although this is a significant aspect, it also lies in the potential it offers to bring research content from libraries to researchers world wide. Potentially, it extends the reach of the collections of both national and research libraries beyond the boundaries of their established research communities and regions. It exploits the collective reputation of libraries as trusted providers of quality information and good metadata. It is a well established fact that libraries are positively associated with books.¹⁷ Providing the full text content of digitised book collections alongside other digital content such as images, videos and audio files, not to mention scholarly content such as articles and theses, means that researchers can obtain richer search

¹⁴ See <http://www.theeuropeanlibrary.org>.

¹⁵ "Report on digital preservation practice and plans amongst LIBER members with recommendations for practical action," Europeana Travel, 2 August 2010, accessed August 25, 2012, http://www.europeana-travel.eu/downloads/D1.3_ET_report_final_23092010.pdf.

¹⁶ See <http://www.europeana-libraries.eu/>.

¹⁷ Cathy De Rosa, Joanne Cantrell, Diane Cellentani, Janet Hawk, Lillie Jenkins, and Alane Wilson, *Perceptions of Libraries and Information Resources: A Report to the OCLC Membership* (Dublin, Ohio: OCLC, 2005), accessed July 5 2012, <http://www.oclc.org/reports/2005perceptions.htm>.

results. Through augmenting the visibility of such content in this way libraries can increase the impact of the significant investment they make in digitisation.¹⁸

Once the project ceases at the end of 2012, the aggregation service will be opened up to all research libraries in Europe. Much work¹⁹ has been done to engage the libraries within the CERL, CENL and LIBER network to define the value propositions of the aggregation service.

The value propositions for libraries include:

- Widened access to library resources via a dedicated portal (run by libraries) for researchers of library content and catalogues
- A one-stop-shop for data processing, comprising a content ingestion workflow that is seamless, automated and extremely fast, together with data enrichment services
- Networking and knowledge sharing

As the service is opened up to more libraries, the sharing of best practice amongst the network and continued development of standards will ensure that content and metadata remains at the highest standard and will support libraries in working more efficiently in the areas of digitisation and metadata.

5. Best Practice & Innovation

Sustainability is not ensured by merely opening up the aggregation service to more libraries. For The European Library there is a pressure to be increasingly innovation and progressive in developing the service in line with technological developments and changes in the behaviour of researchers. For libraries there is pressure to constantly increase the visibility, impact and accessibility of their collections. Collaboration and exploitation of the library networks enables libraries to contribute to the development of, and have access to, new digitisation technologies and best practice. This is exactly what Europeana Newspapers²⁰ facilitates. The Europeana Newspapers aims at the aggregation and refinement of newspapers content, which will be made available through The European Library and Europeana. In addition it addresses challenges particularly associated with digitised newspapers. Through the project, 13 research and national libraries will make digitised newspaper content (from the early 19th century) more available and accessible than it has even been before.

By bringing a mass of newspapers content from all over Europe together in this way Europeana Newspapers achieves the following:

- Contributes to creating a critical mass of content on the First World War and providing an invaluable resource to researchers and citizens alike by making newspapers published during this period available via Europeana;

¹⁸ European Commission, Maurice Lévy, Elisabeth Niggemann, and Jacques De Decker. *The New Renaissance: Report of the Comité des Sages* (Brussels: European Commission, 2011), http://ec.europa.eu/culture/documents/report_comite_des_sages.pdf.

¹⁹ Susan K. Reilly, Marian Lefferts, and Martin Moyle, "Collaboration to Build a Meaningful Connection Between Library Content and the Researcher," *New Review of Information Networking* 17, no. 1 (2012), doi:10.1080/13614576.2012.678139.

²⁰ <http://www.europeana-newspapers.eu/>.

- Adds new value and meaning to newspapers holdings by showing individual newspaper collections in a broader context;
- Creates new search and display capabilities through the development and application of new technologies.

There are other benefits to the work of this project as the procedures with which newspaper content will be upgraded include Optical Character Recognition (OCR), Optical Layout Recognition (OLR)/article tracking, Named Entity Recognition (NER), and page class recognition. OCR will allow users to search the full text of the newspapers. OLR will facilitate searching by article title or section and, the application of named entity recognition means that users can find content related to places, people or things. In other words, the digitised content can be searched more easily, deeply, efficiently, and in such a way as to unveil new relationships and contexts. For each of these technical tasks best practice recommendations will be identified and published. This will be of huge benefit to the broader network of libraries with the CERL, CENL and LIBER networks. It will help reduce the cost of newspaper digitisation projects and increase the accessibility of digital newspaper collections now and into the future.

6. Challenges

As was first recognised in Europeana Travel, digitisation necessitates investment in digital preservation. As libraries step up their digitisation activity the challenge of digital preservation increases. A common thread between all of LIBERs digitisation projects is that they highlight the need for libraries to engage with, and work in a collaborative way to address the challenges related to the preservation of digital cultural heritage.

So how can libraries collaborate to address some of the challenges related to the preservation of digital cultural heritage?

One way is to engage with the digital preservation community as a whole. Through the Alliance for Permanent Access to the Record of Science in Europe Network (APARSEN)²¹ project LIBER libraries have the opportunity to engage in a network of excellence for digital preservation and to contribute to a common vision. Collaboration with a diverse set of practitioners, including both public and private stakeholders, means that libraries can stake their place in the common vision for digital preservation and ensure that the issues surrounding the preservation of digital cultural heritage are represented in this vision.

A more unified vision also means a stronger business case for investment in digital preservation. Taking the findings from the Blue Ribbon Task Force,²² which recommended the definition of roles and responsibilities among stakeholders to ensure an ongoing and efficient flow of resources to preservation throughout the digital lifecycle, APARSEN partners are working together to assess how prepared we are in terms of economically sustainable preservation and how the following conditions for economically sustainable digital preservation can be created:

- Recognition of the benefits of digital preservation on the part of key decision-makers;
- Incentives for the decision-makers to act in the public interest;
- A process for selecting digital materials for long-term preservation;

²¹ <http://www.alliancepermanentaccess.org/index.php/aparsen/>.

²² Blue Ribbon Task Force, "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information." 2010, accessed August 27, 2012, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

- Mechanisms to secure an ongoing, efficient allocation of resources to digital preservation activities;
- Appropriate governance of digital preservation activities.

Participation in a network of excellence also facilitates the exchange of experience and expertise. In the case of APARSEN, this means that gaps in training are identified and filled and expertise is exchanged via various mechanisms such as multistakeholder workshops, Webinars and roadmaps. One of the roadmaps to be produced will be aimed at national and research libraries. Based on a survey of LIBER members, it will document the current situation, identify best practices and reveal the common digital preservation challenges that need to be addressed by libraries going forward.

The intended result of the APARSEN project is a stronger and wider body of expertise, but also a strengthening of, and a deeper understanding across, the network of stakeholders.

7. Conclusion

The push for the digitisation of cultural heritage is both a challenge and an opportunity for libraries in Europe. Digitisation is costly and libraries must find ways of ensuring that it is carried out in the most cost effective and efficient manner. One of the best ways to justify the expenditure of public money on such activity is to make the digitised content as accessible and visible as possible by applying best practice in metadata and pushing content out through an aggregator. The application of new techniques and technologies can add value to the content by making it more searchable, providing context and ultimately making it a useful resource for research and innovation.

Digital preservation is also a costly activity and it is up to libraries as custodians of digital cultural heritage to make the case for it and to engage in best practice.

Collaboration and the use of networks presents the opportunity to for libraries to work more effectively and efficiently. For the LIBER libraries, collaboration has facilitated the development of best practice, enabled European libraries to be leaders in the use of digitisation technologies such as OCR, and given libraries the collective voice to engage in the wider dialogue surrounding digital preservation.

Sustainability is at the heart of digitisation, both in terms of accessibility and preservation. The next step for LIBER is to engage in the development of a sustainable European infrastructure for digitised cultural heritage (Europeana Cloud). The creation of a critical mass of high quality digitised content means that content providers and aggregators, across the European information landscape, urgently need a cheaper, more sustainable technical infrastructure that is capable of storing both metadata and content. Researchers require a digital space where they can undertake innovative exploration and analysis of Europe's digitised content. Europeana Cloud will create a cloud-based infrastructure capable of delivering cost-efficient content and metadata storage for stakeholders across Europe. It brings libraries and other content providers together with other stakeholders, such as technical solutions provider and researchers to provide a trusted, efficient, cheaper and more effective way of making Europe's cultural heritage accessible.

Collaboration and engagement in the development of best practice is the key to discovering sustainable solutions to the challenges of digitisation and digital preservation. It encourages innovation and increases competitiveness. It can also lead to further collaboration and help to forge connections with new communities.

Experiences from Digidaily

Inter-Agency Mass Digitization of Newspapers in Sweden

Heidi Rosen, Torsten Johansson, Mikael Andersson and Henrik Johansson

Abstract

The project Digidaily is a development project and collaboration across authority borders, in which the National Archives of Sweden and the National Library of Sweden, “KB”, are developing rational methods and processes for digitising newspapers. Once the project is completed, we are hoping to transfer to a permanent operation and start digitising our entire collection of 122 million newspaper pages. Digitising cultural heritage is currently a topical issue for many cultural institutions. Many countries began this work several years ago, but the large amount of material usually means that for reasons of costs and handling, only parts of collections or small amounts can be digitized. These are some of the reasons why we at the National Library of Sweden have waited until now with our digitising.

Authors

Heidi Rosen has studied Graphic Arts Technology at KTH, the Royal Institute of Technology in Stockholm. She is currently working as a project manager at the Newspaper Division at the National Library of Sweden, where she is responsible for the library’s part of Project Digidaily.

Torsten Johansson has studied paper conservation at The Royal Academy of Fine Arts in Copenhagen. He has been working with film preservation at the Swedish Film Institute, as a photo conservator at Stockholm City Museum and as division manager for the reproduction unit at The National Library of Sweden. He is currently working as a Division manager at the Newspaper Division at the National Library of Sweden, the Digidaily project is an important part of the units’ future.

Mikael Andersson is currently working as a project manager of Project Digidaily. He is also Production Manager at the National Archive of Sweden department MKC

Henrik Johansson received both his MS in Engineering Physics (2003) and his Ph.D. in Scientific Computing (2009) from Uppsala University, Sweden. He has been working as a Production Developer at the Digital Production Division at the National Library of Sweden where he was responsible for the library’s digitization development.

1. Background

KB has during the years both managed and participated in several projects relating to digitization of large volumes of newspapers. KB soon found that mass digitization of newspapers did not fit within KB’s walls, either physically or organisationally. At the same time, discussions began with the National Archives, which had set up a digitization factory, Media Conversion Centre, “MKC”, in Fränsta in Sweden, mainly to digitize church records. The operation is the largest of its kind in Europe, and the capacity is around 100 000 scanned images per 24 hours. The discussions formed the basis for an application to the Swedish Agency for Economic and Regional Growth for funding from the EU’s structural funds. The application was approved and the project Digidaily started in April 2010.

2. The Collaboration

Our experience to date from the collaboration between our two public authorities has been positive. Cultural differences and the physical distance (450km) between KB and the MKC are, however, aspects that must be considered. In the project Digidaily, we have worked hard to get closer to each other, and we try to meet once a month for joint project meetings. We have also tried letting staff from each authority work at the other authority, with very good results. We carry out study visits together and in between times we keep in contact by telephone, email and meetings held via Skype. But meeting in person is quite clearly the most effective and rewarding way. In other words, a generous travel budget is of significance for a well-functioning project run at a distance.

An important part of the project is to share information. The project has a common online platform for sharing information and documents called Projectplace. In this way, every participant in the project can stay up to date and read memos, time plans, requirement specifications and other important project documents. Projectplace also provides an opportunity to share desks during telephone conferences and, for example, review production and time plans.

Cultural differences are more difficult to overcome. KB is an academic public authority, which often works in a project format, while MKC is a highly efficient production unit, so collisions of culture do occur. But, meeting often and discussing can prevent misunderstandings.

In summary, it could be said that there are lots of positive aspects of working in a development project with another public authority. We have had time to work out and discuss a model that suits both authorities. The wish to maintain high quality in combination with keeping costs down permeates both authorities' attitude to the project, which is an important starting point for a successful collaboration.

3. The Collection

Swedish newspaper publishers deliver three legal deposit copies of all Swedish newspapers printed. One copy stays at KB, one goes to Lund University Libraries and one goes to a company called A2D for microfilming. KB has 31 600 meters of newspapers or approx. 122 million pages out of these are 70 million pages on microfilm.

The collection consists of the so-called official national copies, which are to be preserved "forever". There is also a large collection of duplicates, and it is mainly these that will be used for digitization. KB is taking the opportunity to take stock of and consolidate the collections, so the duplicates will afterwards be destroyed to give space for new incoming newspapers. In those cases where the official national copies are in a poor state, the duplicate will replace or supplement the torn national copy and will therefore be kept.

The unique aspect of KB's collection is the large number of duplicates, which distinguishes the collection from many other library collections around the world. But having more than one copy to consider poses challenges to the project and raises a lot of questions. For instance: When should one mend an existing torn newspaper? When should a supplementary copy be looked for? How much time should be spent searching for alternative material? How should the handling of defects/issues/pages be set up between MKC and KB? How should supplementary material be handled in the metadata (file naming, etc.)? What is the borderline for rejection, how much can be allowed to be torn, what shall KB and its end users accept?

But the benefits outweigh. The use of duplicates permits more efficient procedures, as the preservation aspect does not need to be considered when handling the material. For example, bound material can be cut open and separated. Scanner types that are not very delicate in their handling can be used, etc. And because most of the material is destroyed, the cost of return freight is lower. In the end, the use of duplicates will be noticeable in the overall price. In those cases the official national copies need to be used, KB now considers if it is possible to allow a “gentle dismantling” of the bound material. This would allow a more rational production and therefore a lower price. Strict rules when dismantling isn’t allowed must however be followed.

4. Requirement Specification

As the project is a development project, changes to the requirement specification during the course of the project are permitted. However, now that we are more than halfway through the project, any changes must be of such a nature that they entail significant improvement to the project in order to be taken into account. Too many changes, or large-scale changes, would have a negative effect on the project and the time plan for the project would be greatly disrupted.

For an example KB chose to change a major requirement a year into the project. KB initially chose to save both an archive and a display file, both in grey scale. After a lot of considering, KB changed its mind, and chose to save only one file, an archive file. The amount of data KB saves this way means that the newspaper pages now can be scanned in colour (8 bits/channel) instead and saved in jpeg2000. Saving all images in colour provides great added value for end users as for example most supplements are colour publications.

In short, the end product is a colour page, with segmentation at article level. Manual segmentation or correction of automatic segmentation will **not** be carried out, as the project is striving to use processes that are as automated as possible. Using rules, the CCM software can be adjusted to suit the specific newspaper it is segmenting. To a large extent, it is the skill of the operator that determines how accurate the segmentation is in the end. Correction of for example headlines and other text blocks will not be done either. Here as well we strive to have an accurate and automated workflow.

The requirement specification states that the file shall be at most 300ppi, unless this has a negative effect on readability. KB wants the end user to be able to print out a page of acceptable quality. The general view is also that the resolution should be around 300ppi in order to get an optimal OCR result. MKC has commissioned Mid Sweden University in Sundsvall to look at, and document how resolution and image manipulation impact on the OCR result.

The files are saved in jpeg2000 according to a KB-specific specification. On behalf of KB, Karl-Magnus Drake at the National Archives has investigated the jpeg2000 standard’s fulfilment of criteria for the static image format for long-term storage.¹

Regarding the metadata, a Swedish METS profile has been created. The METS profile is a result of collaboration between the National Library of Sweden, the National Archives of Sweden and other Swedish archives. The basis is a choice of metadata standards such as METS, MODS, PREMIS, MIX and ALTO. This METS profile can also be useful for other digitization projects within the cultural heritage sector.

¹ jpeg2000 – utredningsrapport [Investigative Report] version 2011-03-24 komplett av Karl-Magnus Drake, Riksarkivet.

5. Overall Planning

In the Digidaily project, we are mainly working with two well-known Swedish newspaper titles, Aftonbladet (1830–) and Svenska Dagbladet (1884–). The newspapers belong to Schibsted Media Group, which is also co-financing the project.

If, at the end of the project, there is any spare capacity, the project groups from KB and MKC will discuss the matter and then decide on a suitable newspaper title that will suit both KB's needs and the MKC's production.

KB will take into account the state, size and volume of the newspaper. Whether it uses antique or Gothic font type, if it has been microfilmed, the legal rights of the newspaper, scientific interest and more. In addition, the project team tries to choose materials that are consistent with MKC's wishes in terms of categories of material:

- Category 1 - bound, torn, fragile paper, the biggest format size.
- Category 2 - bound, where most can be taken apart and only a few are kept still bound, fair paper quality.
- Category 3 - tabloids stapled but not bound.
- Category 4 - Official National Copies

Delivery plans are worked out between KB and MKC in order to fulfil to the needs of both organisations as much as possible.

6. Production

6.1 Workflow system

The workflow system is the unifying tool that supports and directs production and processes within Digidaily; the workflow system could be called the spine of the project. The workflow system is constructed in modular form and is developed by a local team of developers at MKC. The process flow is a sequential flow, with status changes that drive the flow onwards.

The workflow system has the following functions:

- To be a database for information about the bundles, issues and pages of the material.
- To add metadata during the course of the production.
- To keep track of and initiate the next process in the flow with the aid of status codes.
- To collect data about the production and create documentation for planning and follow-up.

Also KB has modules for its part of the operation. The material is registered already at KB with basic data, which then will follow the newspaper until the digitization is complete. KB makes an export from its newspaper database, which is entered into the workflow system with basic information about the name of the newspaper, the start and end dates of the bundles and comments on supplements, editions, condition, etc.

Both MKC and KB will be able to enter the workflow system and trace the progress of the material, see the image files, extract statistics, etc.

6.2 Delivery

Using an annual delivery plan as the basis, MKC collects material from KB in Bålsta outside Stockholm. Special transport boxes have been developed for the transport of sensitive materials, such as official national copies. For material that is slightly tougher, KB's ordinary transport trolleys are used.

Each case is followed by a printed packing list from the Workflow system. The official national copies have further identification, with documentation and receipts to safeguard their controlled return.

6.3 Archiving

The material collected is set up in MKC's incoming archive and the archive location is registered in the workflow system. Official national copies are kept in the transportation cases in the incoming archive to minimize handling of the material and ensure secure archival keeping.

6.4 Preparation

The preparation process consists of two sub-processes—*Go through* and *Take apart*. The process has two purposes. One is to capture and record metadata in the bundle, issue and page, and the second is to prepare the material for an efficient image capture.

6.4.1 Go through

The operator goes through the bundle issue-by-issue, page-by-page, and assesses the condition of each individual page. Three levels of divergent condition can be registered in order to communicate to KB that the condition of the page will affect the end result.

- **Level 1** - Lightly damaged. Pale printing, small areas of loss, impact and/or small marks/stains in limited areas. The context of the article can be understood.
- **Level 2** - Severely damaged. Areas that cannot be read even with the eye, and/or parts missing from the page, so that the article cannot be understood.
- **Level 3** - No original or the whole or at least half of a page is missing.

KB will receive reports of the level of rejects via the workflow system and KB can then decide whether or not to search for any better copies. In order for the flow of the material not to be disrupted, the damaged copies continue in the production chain. If a better page/issue is delivered from KB, it is scanned and then replaces the less good page/issue.

The operator also decides which scanning line type is appropriate for the material and checks the pre-registered information in the Workflow system in terms of date and number of the issue, and how many pages each part contains.

The operator completes the information in the workflow system with information about, for example:

- Name of supplement and/or section
- The genre it belongs to—Supplement, news bill or section
- The edition of the issue.
- The number of the issue.

Once the bundle has been gone through, an assessment is made whether it is suitable to separate the bundle into loose pages. Destruction or return copies can be taken apart if the text information can be assured afterwards.

6.4.2 Take apart

The operator takes apart the bundles with great caution. Knives are used to divide the bundle into smaller piles, and then an electrical cutter is used to get an even and smooth cut surface.

6.5 Image capture

For the moment there are two methods of capturing images—book scanning and wide format sheet fed duplex scanning

For book scanning, the scanner models Zeutschel OS 14 000 A1 and A0 are used. The model allows image capture of two pages at the same time, which can be divided into separate images. For wide format sheet feed scanning, the scanner model SUPAG Mediascan 880c is used. The model allows image capture of double-sided pages and the entire spread. Front and back pages are scanned in one feeding. Spreads are being divided into four separate images. A function to number and organize files on 4-page scanning is also available.

6.6 Technical quality control

In order to safeguard the quality of the file, a fully automated technical control is carried out on all files. From that process metadata are lifted out of the data file and recorded in the Workflow system as a basis for METS.

In case of deviation from the established quality requirements, the file will be returned to the image capture process for re-scanning.

A performance file will be saved in the final package in order to guarantee quality. The performance file includes the latest measurement data from the quality measurement of the image capture equipment. Software manufactured by KB, Colorite,² will be used for this purpose.

6.7 Creating jpeg

In order to ensure convenient handling, MKC creates jpeg copies of all files.

- **High-resolution jpeg**
A high-resolution jpeg copy is created for the OCR software.
- **Low-resolution jpeg**
A low-resolution JPEG file is created for use in the ocular quality control process.

² Henrik Johansson, "Colorite: A Flexible Cross-Platform Software Solution for Automatic Image Quality Analysis Using Arbitrary Targets," in *Archiving 2011: Final Program and Proceedings*, vol. 7 (Salt Lake City, UT: IS&T, 2011), 199-204.

6.8 Ocular quality control

An ocular quality control is carried out to ensure the image capture has been satisfactory. Today the operator examines 100 per cent of the images visually. A future development of the ocular control aims to use statistically methods to extract significant number of images that will be examined visually. If the quality control shows that the quality of the image does not meet established quality levels, the image is reported for re-scanning.

6.9 Approval of image capture

The decision point for approving an issue initiates the start of two flows; the flow of the digital file and the continued flow of the physical issue.

For the digital workflow the creation of a *Jpg2000* file is initiated. For the physical issue flow the delivery process for the *Return* copy and *official national copy* is initiated.

6.10 File flow

KB's archive file shall be of format JP2000. The JP2000 copy is created from TIFF, according to specifications from KB

6.11 OCR interpretation process

The OCR result is the primary result of the digitization. The resolution of the image is based on the quality of the OCR result. Testing of how the resolution impacts on the OCR result is being carried out by Mid Sweden University on behalf of the project. The project has not yet found an adequate way of setting requirements for the correctness of the OCR result. The content conversion software used for the segmentation and OCR interpretation comes from the Norwegian company Zissor.

The OCR process consists of three subsidiary processes—*OCR*, *quality control* and *export*. The Workflow system creates a log file that controls the layout analysis set up of the different materials.

- **OCR**
According to the layout analysis set up, the images are being interpreted. No manual article segmentation is made in project Digidaily.
- **Quality control**
An audit is made of the OCR result for a statistical sample of the images.
- **Export**
Once a satisfactory quality level has been achieved in the interpretation, ALTO files are exported from the program.

6.12 Creating METS

The METS file for each issue is created from the data collected about the material in the Workflow system. As mentioned before, KB has clearly specified XML-schema for the METS-file.

6.13 Packaging deliveries

MKC creates a SIP- package, one for each issue, containing all the object files and a METS file with metadata about the content in the package. Each packet should contain the following parts:

- jpeg2000/page
- METS/issue
- ALTO/page
- Performance file/issue

6.14 Delivering packets

The completed packages are sent via ftp to KB's storage system. A delivery receipt with information if the delivery is approved or disapproved is sent back to MKC.

6.15 Flow of the physical issue after digital delivery

Material to be returned is sent back to KB for further handling and archiving, and material not to be returned will be destroyed as instructed by KB.

6.15.1 Re-delivery to KB

For the copies that are to be returned to the KB, the period of retention at MKC should be kept as short as possible. Material that has been taken apart are packed into boxes before labelling and delivery. A pack-list is being extracted from the Workflow system and delivery notes are attached to the packages. Official national copies have additional registration and receipts to ensure the delivery procedures.

According to the established delivery schedule, material is returned to KB for further registration and archiving. The re-delivery to KB initiates the start to reform and strengthens the collection. Better copies will either replace damaged official national copies or the decision will be made to keep both if they are in a very poor state. The material will be stored for final archiving.

6.15.2 Destruction Process at MKC

Once delivery of digital images has been approved the destruction process begins. To ensure that the right material is sent for destruction, strict rules and protocols must be followed. A recycling company will collect the material for further destruction.

7. Summary

The strength and success of the project lies in KB's and MKC's project groups being focused and having the same objective: ensuring that quality can be allied to a competitive price. By working together, we can also benefit from the joint competences of the staff in an effective way. The chance of trying it out, in terms of both technology and procedures, has also resulted in an efficient workflow. And lastly, but not least, the spine of the entire project, the workflow system, which makes it possible for both MKC and KB to keep track of every single page.

As an experienced production unit, MKC has detailed knowledge about all the subsidiary costs of the flow, which gives us a good tool for further efficiencies. For example, currently the average cost for preparation absorbs around 47% of total costs, and scanning 39%. The better quality of the material, the lower preparations costs and therefore also a lower total cost.

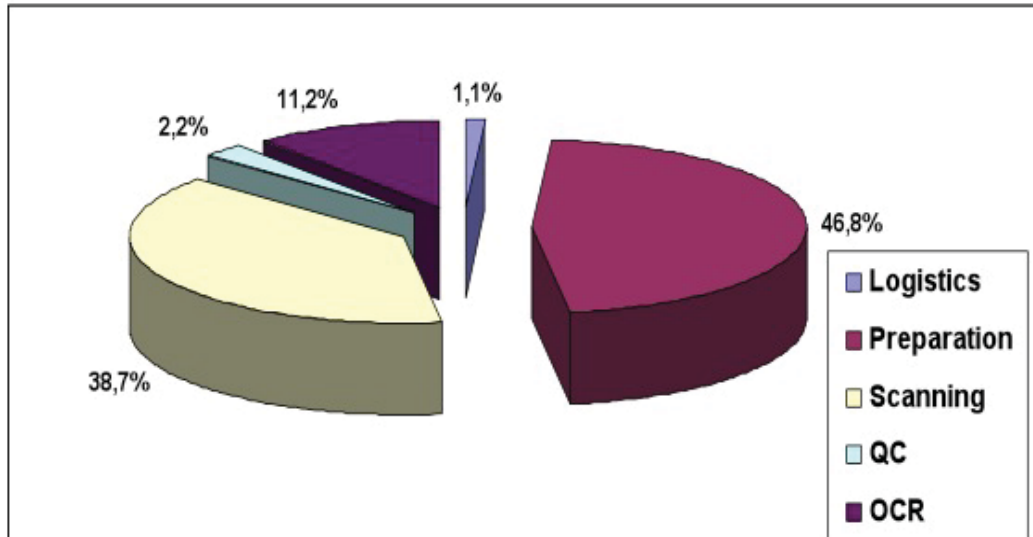


Figure 1. Relative shares of the cost

Depending on the size and condition of the newspaper material, the cost of a digitized page, including OCR, is today €0.25/page (Category 3 - stapled tabloids) up to €0.93/page (Category 1 - bound, torn, fragile paper, up to A0). In US dollars the price span is around \$0.31 – \$1.41/page inc. OCR.

For further information, please visit our blog Digidaily: <http://digidaily.kb.se/>

Preserving Images: What Do We Need to Know?

Chemins de la mémoire

Les archives audiovisuelles au secours de l'identité d'une organisation internationale africaine

Adama Aly Pam

Archiviste paléographe

Résumé

Trente ans après le transfert de son Siège de Paris à Dakar, la Banque Centrale des États de l'Afrique de l'Ouest (BCEAO) décide d'organiser ses archives. Cette préoccupation tardive, mais salutaire découle d'une prise de conscience d'une génération de banquiers centraux soucieuse de célébrer une histoire. En effet, le transfert du Siège est accompagné d'une africanisation du personnel perçue comme un acte d'indépendance. Le départ à la retraite d'une bonne partie des acteurs de cette africanisation a été le déclic d'une forte demande de mémoire et de célébration. C'est ainsi que la Banque s'est lancée dans une vaste politique d'exhumation et d'organisation de la mémoire. Cela s'est traduit par la rédaction d'une somme monumentale de l'histoire institutionnelle de l'Union monétaire ouest-africaine, la création d'un musée de la monnaie et l'organisation des archives et de la documentation à l'échelle des 23 sites de la Banque. Le programme de gestion des archives audiovisuelles est une composante essentielle de l'organisation de la mémoire institutionnelle. Il vise à mettre en place une politique de traitement de ces documents, suivant trois axes majeurs : la sauvegarde du patrimoine audiovisuel, l'amélioration de sa gestion et sa valorisation. Le présent exposé restitue le cadre méthodologique du dispositif de gestion des archives audiovisuelles que nous avons mis en œuvre.

Auteur

Archiviste paléographe, monsieur Pam est docteur en Histoire de l'université Cheikh Anta Diop de Dakar. Il commence sa carrière aux Archives nationales du Sénégal avant de rejoindre la Banque Centrale des États de l'Afrique de l'Ouest où il conduit avec une équipe d'archivistes un important programme de modernisation des archives et de la Documentation de l'Organisation régionale. Professeur vacataire à l'Université Cheikh Anta Diop, il a participé à la mise en place du programme pédagogique de l'enseignement à distance d'archivistes d'entreprise. De 2007 à 2010, monsieur Pam a été élu Président de l'Association sénégalaise des bibliothécaires, archivistes et documentalistes. Il est auteur de plusieurs rapports et études sur les questions touchant les archives et les bibliothèques au Sénégal.

1. État des lieux des archives audiovisuelles de la BCEAO

1.1. Inventaire des archives audiovisuelles

Le recensement des archives audiovisuelles a été réalisé au Siège, dans les Directions nationales et au Secrétariat général de la Commission bancaire au cours de la période du 8 novembre au 29 décembre 2006.

À cet effet, quatre formulaires ont été élaborés pour recenser les archives audiovisuelles par type de support, notamment les photographies, les enregistrements vidéo, les enregistrements audio, et les affiches. Le recensement a permis de dénombrer et de faire ressortir la nature des documents, leurs supports et leur état de conservation.

1.1.1. Volume du fonds

Il ressort des résultats du recensement que la Banque dispose d'une collection d'archives audiovisuelles composée d'environ 75,386 unités. Le dépouillement met en évidence une répartition géographique des archives audiovisuelles de la Banque, marquée par une forte concentration au niveau du Siège, qui totalise 44,002 documents audiovisuels, constituant environ 58 % du fonds total. Les Directions nationales disposent d'un fonds d'archives audiovisuelles compris entre 7,92 % (DN Sénégal) et 0,15 % (DN Togo) du fonds total.

Recensement des archives audiovisuelles de la BCEAO : répartition thématique par site

	Siège	SGCB	DN Bénin	DN B Faso	DN RCI	DN G Bissau	DN Mali	DN Niger	DN Sénégal	DN Togo	Total	%
Construction & inauguration des immeubles abritant les sites de la Banque	9 843	37	744	271	141	192	1 179	0	934	0	13 341	17
Emission, nouvelles gammes de billets, démonétisation	561	1	0	244	7	8	18	18	402	19	1 278	2
Cérémonies protocolaires	1 802	392	263	142	0	0	0	1	540	0	3 140	4
Réunions statutaires de l'UMOA et de l'UEMOA	7 208	0	183	209	220	0	1 038	137	866	3	9 864	13
Autres réunions	2 172	384	0	0	0	242	0	0	59	5	2 862	4
Activités du Gouverneur : audiences avec les chefs d'Etat, parrainages, rencontres internationales,	1 705	188	2	14	317	96	82	13	79	6	2 502	3
Fêtes commémoratives	2 653	0	415	1	291	1 245	2	8	472	69	5 156	7
Cérémonies sociales : arbres de Noël, sorties récréatives, sport, départ à la retraite, colonies de vacances, fêtes diverses	5 413	870	2 627	2 531	1 131	850	2 255	4 192	2 687	0	22 556	30
Formations, séminaires, colloques	3 620	85	0	0	4	0	359	59	5	7	4 139	5
Autres	9 025	0	264	1 039	1 151	0	2	0	0	2	11 483	15
Total	44 002	1 957	4 498	4 451	3 262	2 633	4 935	4 428	6 044	111	76 321	100
%	57,65	2,56	5,89	5,83	4,27	3,45	6,47	5,80	7,92	0,15		100

Source : BCEAO

Par type de document, les images photographiques (75,048 unités) représentent 99 % du fonds d'archives audiovisuelles de la Banque. Les enregistrements vidéo et audio comptent respectivement 286 et 52 unités.

Recensement des archives audiovisuelles de la BCEAO : répartition thématique par type de document

	Photographies		Enregistrements vidéo	Enregistrements audio	Affiches	Ratio photos	Ratio Vidéo	Ratio audio	Ratio Affiches
	Nbre d'albums	Nbre de photos							
Construction & inauguration des immeubles abritant les sites de la Banque	130	13 333	5	0	0	17,78	1,75	0	0
Emission, nouvelles gammes de billets, démonétisation	12	1 063	78	48	8	1,42	27,27	92,31	72,73
Cérémonies protocolaires	30	3 134	4	0	0	4,18	1,4	0	0
Réunions statutaires de l'UMOA et de l'UEMOA	86	9 843	12	1	0	13,13	4,2	1,92	0
Autres réunions	23	2 160	5	0	0	2,88	1,75		0
Activités du Gouverneur : audiences avec les chefs d'Etat, parrainages, rencontres internationales,	36	2 455	20	0	0	3,27	6,99	0	0
Fêtes commémoratives	46	5 128	15	0	0	6,75	5,24	0	0
Cérémonies sociales : arbres de Noël, sorties récréatives, sport, départ à la retraite, colonies de vacances, fêtes diverses	303	22 452	51	1	0	29,94	17,83	1,92	0
Formations, séminaires, colloques	37	4 032	87	0	0	5,38	30,42	0	0
Autres	111	11 448	9	2	3	15,27	3,15	3,85	27,27
Total	814	75 048	286	52	11	100	100	100	100

Source : BCEAO

1.1.2. Composition thématique du fonds

L'analyse thématique du fonds révèle qu'il est composé en majorité (30 %) de documents relatifs aux activités sociales (sorties récréatives, colonies de vacances, activités sportives, arbres-de-Noël, départs à la retraite). Les documents relatifs aux programmes de construction ou d'inauguration des sites de la Banque entre 1958 et 2000 comptent pour 17 % du fonds, et les réunions des instances de l'UMOA pour 13 %. Les archives relatives à l'émission et à la démonétisation représentent environ 2 % du fonds, et celles liées aux activités du Gouverneur (audiences avec les Chefs d'État, rencontres internationales, parrainages) comptent pour 3 %. Les activités de formation constituent 5 % du fonds total.

1.2. État de conservation des archives audiovisuelles

Dans plusieurs sites, les conditions de conservation des archives audiovisuelles ne sont pas conformes aux normes requises en la matière. Les documents sont conservés dans les bureaux et dans quelques salles de prétraitement, alors qu'ils devraient être stockés dans des espaces de conservation spécifiques. Ces espaces doivent être protégés de la poussière, de la lumière et mis hors de proximité des transformateurs et moteurs électriques. La température des locaux doit se situer de manière constante autour de 20 °C et d'une humidité relative de 40 %. Les supports doivent être conditionnés dans des boîtes qui évitent toute déformation ou rayure de la couche polycarbonate, en ce qui concerne les CD contenant des données numériques.

1.3. Inexistence d'instruments de recherche

Il n'existe aucun instrument de recherche des archives audiovisuelles de la Banque. Cette situation est d'autant plus complexe que certains de ces documents nécessitent le recours à un appareil de lecture pour accéder au contenu (cassettes vidéo et cassettes audio). Par ailleurs, les documents photographiques sont souvent sans indications sur l'identité des personnes, les lieux, la date et le contexte ou l'objet.

2. Plan de gestion des archives audiovisuelles

Au regard de l'importance du fonds des archives audiovisuelles de la Banque, et pour pallier les insuffisances constatées lors du recensement, il est proposé de mettre en place un programme global visant à assurer l'uniformité de leur gestion sur tous les sites de la Banque, dans le respect des normes internationales en vigueur en la matière.

Ce programme comprend l'identification des archives audiovisuelles et la mise en place d'outils adéquats pour leur gestion. Il inclut également l'application des normes idoines de conservation des documents sur tous les sites de la Banque, ainsi que la mise en œuvre d'un plan de numérisation et de valorisation du fonds.

2.1. Identification des archives audiovisuelles

Le programme s'attellera à mettre en place un dispositif d'identification des archives audiovisuelles, en ayant recours aux agents de la Banque ou à d'autres témoins. En effet, l'identification d'un document constitue une information très utile au processus d'évaluation et de tri. Il consiste à établir une carte d'identité précise de chaque document afin de permettre son identification de manière univoque et précise. Les règles de description des documents d'archives obligent à rechercher et à consigner dans une base de

données son titre, son auteur, sa date de création et son contenu. Pour ce faire, une équipe constituée d'archivistes, d'historiens et d'agents désignés en raison de leur connaissance de l'histoire contemporaine de la Banque sera mise en place au Siège, et en cas de besoin, dans chaque Direction nationale.

2.1.1. Identification des albums sans légende

Cette opération consistera à identifier les événements auxquels se rapportent les albums ne disposant aucun renseignement permettant leur description. Il s'agira, pour l'archiviste chargé du traitement, de recourir aux personnes ayant participé à l'événement afin de déterminer de manière relative ou absolue la date et le contexte d'élaboration des photographies. Cette opération nécessite de la part de l'archiviste plusieurs recoupements pour valider l'information recueillie auprès des personnes ciblées (en activité ou à la retraite).

2.1.2. Identification des documents vidéo et audio sans légende

Il s'agira de visionner les cassettes vidéo afin de déterminer le contexte et l'objet de la production du document. Le recours aux personnes témoin des événements peut être envisagé. Quant aux documents sonores, l'équipe procédera à l'écoute des documents pour déterminer leur contenu. Il sera procédé à une transcription textuelle des discours pour en faciliter l'identification.

2.2. Outils de gestion

Un plan de classification, un bordereau de saisie et un calendrier de conservation et d'élimination sont élaborés et mis à la disposition des archivistes de la Banque pour le traitement des archives audiovisuelles.

2.2.1. Élaboration d'un plan de classification

La classification a pour but d'identifier et de regrouper selon une logique déterminée les documents d'archives appartenant à un fonds ou à une collection, en vue d'en faciliter le repérage. Dans le cadre des archives photographiques, les fonds peuvent être considérés en fonction des différentes photographies. Cette option peut avoir des limites si on ne parvient pas à déterminer toutes les photographies, notamment pour les documents anciens au regard du fait que les photographies ne sont généralement pas regroupées en agence. La totalité de la collection peut également être envisagée comme une entité archivistique unique. Pour des considérations pratiques, nous choisissons la dernière option et proposons le plan de classification ci-après :

100 — Construction des agences de la BCEAO

- Plans et maquettes
- Travaux de construction
- Inauguration des sites

200 — Rencontres statutaires

- Conférences des Chefs d'État
- Conseils des ministres de l'UMOA
- Conseils d'Administration de la BCEAO

300 — Cérémonies protocolaires

- Prestations de serment des Gouverneurs
- Installations des Directeurs nationaux
- Fêtes commémoratives

400 — Cérémonies sociales

- Arbre de Noël
- Sorties récréatives (colonies de vacances, championnats divers)
- Départs à la retraite

500 — Colloques et séminaires

- Séminaires
- Colloques
- Formations diverses

600 — Campagnes de mise en circulation de billet de banque ou de démonétisation

- Mise en circulation de billets et pièces
- Opérations de démonétisation.

Ce cadre de classification, donné à titre indicatif, indique l'architecture générale à suivre pour la représentation et la description des archives audiovisuelles de la Banque. Il sera adapté pour inclure les spécificités des fonds des différents sites.

2.2.2. Bordereau de saisie des archives audiovisuelles

Le bordereau de saisie des archives audiovisuelles devrait servir pour la description des documents, avant leur entrée dans la banque de données dédiée. Chaque document doit faire l'objet d'une fiche descriptive qui détermine le contenu, les auteurs et l'objet du document. Le détail des champs du bordereau se présente comme suit :

INTITULÉ	CONTENU
LIEN	Lien hypertexte du document s'il est numérisé
COTE	Cote attribuée au document dans le plan de classification
UNI DESC	Unité de description (photographie, album)
TITRE	Mention de l'événement au cours duquel les documents ont été réalisés
TYPE DE SUPPORT	Définir le type du document (photographie, vidéo, affiche, etc.)
AUTEUR	Auteur du document
BIOGR	Déterminer la date de naissance ou de décès de l'auteur, le type de contrat le liant à la Banque afin de s'assurer du statut juridique des documents de l'auteur.
ENTREE	Mode d'entrée du document (don, versement, achat)
ACCES	Si le document est communicable ou non

INTITULÉ	CONTENU
REPRO	Si la reproduction est autorisée ou non en raison des droits d’auteurs ou des conditions physiques du document
DROIT PATRIMONIAL	Le titulaire du droit patrimonial
DROIT MORAL	Le titulaire du droit moral
LANGUE	Langue du document
ORIGINAL	État du document (si original ou copie)
DOC LIE	Il s’agit de signaler les documents liés au document décrit : dossiers d’archives, ouvrages. Un album photo peut renvoyer à une manifestation (Conférence des Chefs d’État) pour laquelle on dispose de dossiers d’archives.
DESCRIPTEURS	Indexation du document — ce champ intègre les indexe géographiques, noms propres, thématiques, etc.
ARCHIVISTE	Nom de l’archiviste auteur de la description

2.2.3. Élaboration d’un calendrier de conservation et d’élimination

Le calendrier doit permettre d’atteindre plusieurs objectifs, à savoir diminuer la masse documentaire à conserver, préserver les documents qui ont une valeur administrative, légale ou historique, réduire le coût de conservation des documents et augmenter l’efficacité de la gestion du fonds. Les règles en vigueur dans le cadre des tableaux de gestion des archives de la BCEAO sont appliquées aux archives audiovisuelles.

2.3. Mise aux normes des conditions de conservation des archives audiovisuelles de la BCEAO

La complexité de l’archivage des documents audiovisuels est liée à l’obsolescence rapide des technologies de lecture des types de documents. La principale difficulté réside dans la disparition des appareils de lecture, dont la plupart ne sont plus fabriqués par l’industrie. Les disques vinyles, les vidéocassettes, les bandes et cassettes magnétiques illustrent bien le cas. De ce fait, la préservation des archives audiovisuelles passe par la conservation des supports et le transfert périodique des enregistrements sur des supports contemporains.

Les normes et standards applicables aux types de documents disponibles à la BCEAO sont présentés dans les sections suivantes. Ils sont notifiés aux structures de la Banque pour prise en compte dans le cadre de la gestion courante des archives audiovisuelles.

2.3.1. Conditions de conservation des documents photographiques

La conservation des documents photographiques doit être conforme aux normes ISO 18902-2001 et ISO 18916-2007, relatives respectivement aux standards de conservation et aux techniques de vérification de la conformité des matériaux (papier, adhésifs, encre), afin de s’assurer qu’ils n’interagiront pas de manière négative sur les documents. Les altérations que subissent les photographies sont d’ordre

biologique (moisissures, insectes), chimique (traces de doigt, adhésifs, pâlissement, affaiblissement) et physique (déchirures, abrasions, plis, craquelures, gondolement).

Pour pallier ces altérations, il est retenu de conserver les archives photographiques de la BCEAO dans des boîtes et pochettes de marque NOMI ou Mylar. Elles présentent l'avantage d'être fabriquées en papier permanent¹ et protègent les photographies contre les interactions négatives liées à l'acidité des supports ordinaires.

Les conditions atmosphériques et hygrométriques à respecter dans les locaux de conservation des documents photographiques sont récapitulées ci-après :

Tableau 1. Prescriptions relatives à l'atmosphère des magasins d'archives photographiques²

Support	Température			Humidité relative		
	Objectif	Fluctuations		Objectif	Fluctuations	
		Quotidienne	Annuelle		Quotidienne	Annuelle
	°C	°C	°C	%	%	%
Négatif noir & blanc	< 18	1	2	30-40	5	10
Épreuve noir & blanc	< 18	1	2	30-40	5	10
Négatif couleur	2	1	2	30-40	5	10
Diapositive	2	1	2	30-40	5	10
Épreuve couleur	2	1	2	30-40	5	10

2.3.2. Conditions de conservation des archives audio et vidéo

Pour les cassettes vidéo et audio, de bonnes conditions d'entreposage et des procédures de manipulation appropriées permettent de prolonger la durée de vie des enregistrements. Par ailleurs, afin de prévenir l'obsolescence technologique de ces types de support, il est proposé de procéder au transfert des données sur des disques compacts CD-R. Les disques compacts sont dupliqués et entreposés dans les conditions conformes aux normes édictées en la matière, présentées en annexes 1 et 2.

2.4. Numérisation des archives audiovisuelles

La numérisation offre une nouvelle façon de conserver l'information et de la rendre plus accessible. Dans le cas des documents audiovisuels, du fait de l'obsolescence rapide des supports ou de leur fragilité, elle

¹ Papier sans acide, antifongique, recommandé pour la confection de boîtes destinées à assurer la conservation à long-terme des archives à forte valeur patrimoniale.

² Dans : Marie-Thérèse Varlamoff, conservation préventive du patrimoine documentaire, document photographiques et film, Paris, Unesco, programme "Mémoire du Monde".

est souvent le seul moyen de sauvegarde. Les fichiers numérisés alliés à des outils de recherche informatiques favorisent une plus grande accessibilité aux documents. La numérisation réduit la manipulation des documents, allégeant ainsi la pression sur les collections. Ces avantages, associés au raccourcissement du temps que le personnel passe à faire des recherches, constituent des arguments économiques de taille qui militent en faveur du choix de cette méthode.

Dans le cadre du programme de gestion des archives audiovisuelles de la BCEAO, les documents à forte valeur patrimoniale sont numérisés en priorité. Les documents dont l'état physique de conservation ne permet pas d'être communiqué aux usagers font également l'objet de numérisation.

2.4.1. Numérisation des photographies

L'organisation du fonds de sauvegarde et du fonds d'utilisation (copie des documents les plus utilisés) doit veiller à assurer la conservation des originaux des documents à forte valeur patrimoniale. La numérisation des photographies à forte valeur ajoutée est effectuée selon deux formats : le format TIFF pour l'archivage et le format JPEG pour la diffusion et l'affichage dans le réseau intranet de la Banque.

2.4.2. Numérisation des documents vidéo analogiques

Conçus pour une durée de 10 ans, les supports magnétiques (bandes et cassettes vidéo) se détériorent vite. Les mauvaises conditions de conservation accélèrent de manière irrémédiable la dégradation de ce type de support. L'humidité, la chaleur et la lumière sont les principales sources de dégradation. Pour prévenir les risques de dégradation et de perte d'informations, nous avons procédé à la numérisation des cassettes et bobines au format MPEG-2, afin de fournir le meilleur compromis entre la qualité de la compression et la qualité de l'image.

2.4.3. Numérisation des documents audio analogiques

En amont de l'opération de numérisation, il convient de procéder à un traitement documentaire qui consiste à faire l'inventaire des documents originaux à numériser:

- identification des contenus grâce aux sources écrites préexistantes, ainsi qu'aux mentions sur les étiquettes; relevés de l'état de conservation des bandes dans la mesure où les dégradations peuvent être constatées à l'œil nu;
- Indexation primaire : saisie des données d'identification dans la base de données documentaire, avant le prélèvement des documents pour l'opération de numérisation. Un numéro **d'identifiant** unique en code à barres sera attribué au document original et porté sur la notice bibliographique. Le système de classification retenu servira de base d'identification des documents. Cette opération permettra d'effectuer un suivi des documents tout au long de la chaîne de traitement, de la numérisation au rangement des originaux dans le magasin de conservation;
- Conservation des lots à numériser : regroupement des unités documentaires dans le but de constituer un corpus cohérent pour la numérisation.

La démarche à adopter pour la numérisation des documents audio est indiquée à l'annexe 2.

3. VALORISATION DES ARCHIVES AUDIOVISUELLES

La valorisation des archives est le but ultime de toute action de conservation. Elle consiste à permettre l'accessibilité des documents et à les faire connaître à travers une série de techniques de communication (expositions, publications scientifiques, brochure de vulgarisation, sites web à thème sur des problématiques spécifiques).

Une politique de valorisation des archives audiovisuelles est mise en place à la BCEAO à partir de l'organisation d'expositions à l'occasion des fêtes commémoratives ou en fonction de l'actualité économique et financière des États de l'Union et par la publication de ces documents dans les rapports et documents officiels de la Banque.

3.1. Organisation d'expositions thématiques virtuelles

Une exposition virtuelle est une présentation exclusivement en ligne d'une thématique documentée à partir des archives de la Banque. Elle peut être composée de documents graphiques, sonores et multimédias. Elle peut être accompagnée de l'édition d'un catalogue électronique sous forme de CD-ROM.

Elle peut être organisée à partir du site Intranet de la Banque à l'usage des agents, ou sur le site internet pour un plus large public. Un comité composé d'historiens, d'archivistes et d'un informaticien pourrait être constitué à cet effet.

3.2. Organisation d'expositions permanentes ou itinérantes retraçant l'histoire de la Banque

La Direction de la recherche et de la statistique pourrait, dans le cadre de la mise en valeur des archives historiques de la Banque, organiser des expositions itinérantes ou permanentes sur des thématiques relevant des missions de l'Institut d'émission. Les expositions permanentes sont organisées sur les différents sites de la Banque, pour servir de lieu de mémoire permettant aux visiteurs ou à tout nouvel agent de la Banque d'avoir un aperçu de l'histoire et du patrimoine de l'institution.

Par ailleurs, l'objectif d'une exposition itinérante est de toucher un public plus large et de renforcer la notoriété de l'institution. Sa réalisation fait généralement appel à une équipe de compétences diverses, notamment la recherche historique, les techniques de conservation, l'édition, le design, l'animation et la promotion.

3.3. Publication de documents à l'occasion des dates commémoratives

La publication de documents à l'occasion des dates commémoratives à l'instar de ceux édités par la Banque à l'occasion du 40^e anniversaire de la BCEAO participe au renforcement de l'identité de l'institution. C'est également un moyen de valoriser le patrimoine documentaire de la Banque.

Annexe 1 : Normes de conservation du disque compact d'archivage de données

Les disques compacts (CD audio, cédéroms et CD-R) sont très fragiles et se rayent facilement. Le moindre défaut ou dépôt de poussière peut constituer un obstacle au faisceau laser et perturber la restitution de l'enregistrement. Par ailleurs, une rayure profonde sur le vernis imprimé peut altérer gravement la couche d'inscription des informations. D'une manière générale, la surface transparente ne devra jamais être touchée à mains nues. Le disque sera maintenu par les bords. Le port de gants en tissu qui ne peluche pas est conseillé lorsque le contact direct avec les faces ne peut être évité. Le marquage des disques est à effectuer à l'aide d'un marqueur-feutre indélébile préconisé par les fabricants. Il est recommandé de limiter les inscriptions à la partie centrale du disque.

Les conditions de stockage tiennent compte de la réaction des disques aux différents facteurs, notamment l'humidité, la lumière et la température. La température dans les magasins de conservation sera maintenue à 20 °C et les fluctuations ne devront pas dépasser 10 °C. L'humidité relative restera comprise entre 20 % et 50 % avec des fluctuations inférieures à 10 %.

Les disques enregistrables une fois (CD-R), qui sont ceux destinés à l'archivage, sont tout particulièrement sensibles à deux facteurs : la température et la lumière. Ils ne devront jamais être exposés aux rayons du soleil.

Le contrôle de l'état des collections est indispensable. L'inspection sur un échantillonnage représentatif devra être effectuée tous les cinq ans. Les moyens de contrôle de la qualité de l'environnement de conservation sont composés de thermo-hygromètre et de thermomètres.

Consignes :

- Ne pas exposer les disques au soleil ou à une lumière forte pendant de longues périodes;
- Manipuler les CD avec précaution et ne pas utiliser de feutre, d'alcool, ni d'étiquettes (la surface supérieure [marquée] du disque est la plus vulnérable, car elle comporte une fine couche de laque pour protéger la surface de gravure);
- Conserver les CD entre 5 °C et 20 °C;
- Le taux d'humidité relative doit être compris entre 20 et 40 %;
- Le gradient de température qui équivaut au choc thermique doit être de 4 °C / heure;
- Le gradient d'humidité relative doit être de 10 % par heure.

Annexe 2 : Numérisation des documents audio analogiques

- Prélèvement des supports originaux;
- Avant la lecture des bandes : rebobinage et rembobinage des originaux sur un magnétophone adapté à cet usage et petite restauration mécanique effectuée sur les bandes le nécessitant;
- Enregistrement et conversion du signal analogique en numérique (sans correction de tonalité) en format Wave (en 48 Khz/24 bits ou en 96 Khz/48 bits);
- Écoute intégrale de l'original et rédaction d'un rapport technique d'écoute pour chaque bande en vue de renseigner la base de données documentaire;
- Sauvegarde du fichier sur le PC de gravure — Gravure d'un CD-ROM pour la conservation (en 48 Khz/24 bits ou en 96 Khz/48 bits) et de trois CD-R (en 44,1 Khz/16 bits) pour la consultation;
- Vérification et contrôle-qualité effectués sur l'ensemble ou sur un échantillon de la production.

Digitization as a Preservation Strategy

Saving and Sharing the American Geographical Society Library's Historic Nitrate Negative Images

Krystyna K. Matusiak¹ and Tamara K. Johnston²

¹*Morgridge College of Education, University of Denver, krystyna.matusiak@du.edu*

²*American Geographical Society Library, University of Wisconsin-Milwaukee Libraries, johnstot@uwm.edu*

Abstract

Digitization as a preservation strategy has been the subject of debate among the members of the cultural heritage community for two decades. The benefits of digitization in expanding access are universally acknowledged, but the recognition of digitization as an option for long-term preservation of analogue materials is still controversial. This paper contributes to this discussion by exploring the ways in which digitization brings a renewed attention to preservation of endangered photographic collections and enables preserving the content of deteriorating visual materials. The authors present the digitization project at the American Geographical Society Library undertaken as an approach to providing access and preserving to over 69,000 nitrate negatives from the historic geographic collections.

Authors

Krystyna K. Matusiak served as a co-investigator of the *Saving and Sharing the AGS Library's Historic Nitrate Negative Images* project. She has been involved in digitization of cultural heritage materials for 11 years. She worked as the Head of the Digitization Unit at the University of Wisconsin-Milwaukee where she planned and designed over 20 distinct digital collections. She also served as a digitization consultant for projects funded by the Endangered Archive Programme and assisted digital library projects at the Press Institute of Mongolia in Ulan Baatar, Mongolia and the Al-Aqsa Mosque Library in East Jerusalem. In conjunction with those projects, she published two papers on using digitization as a preservation strategy in developing countries. Currently, she works as an Assistant Professor at the University of Denver where she teaches classes on digitization in the Library and Information Science program

Tamara K. Johnston served as the coordinator of the project, *Saving and Sharing the AGS Library's Historic Nitrate Negative Images* and also co-authored the grant submissions to the National Endowment for the Humanities. Johnston has worked in museum collection management for the past 20 years, focusing on preservation and intellectual access issues. Currently, she works as a project manager and preservation specialist at the American Geographical Society Library. Previously she spent twelve years at Bryn Mawr College's Art and Archaeology Collection in Bryn Mawr, PA where she was the Collections Manager of "a museum without walls." She also worked with museum data management and preservation projects at the Kohler Foundation, the Penn Museum of Archaeology and Anthropology, the Pennsylvania Academy of Fine Arts Museum, and the Milwaukee Public Museum.

1. Introduction

Digitization as a preservation strategy has been discussed among the members of the cultural heritage community since the early days of digitization projects. The debate has focused on the tension between long-term preservation and extending access. Initially, digitization was accepted as a form of copying for wider and easier access, but not as a method for creating preservation quality copies of original

materials.¹ Digitization for preservation, however, is slowly gaining recognition. It has brought a renewed attention to preservation of endangered archival materials and is often viewed as the best option for capturing and preserving the content of deteriorating film negatives and other non-print resources.²

Archival photographic, audio, and video collections provide rich and often untapped source of historical evidence, but their preservation is problematic due to complex and deteriorating formats, limited documentation, and the lack of reformatting standards.³ Historically, audiovisual resources have been recorded on fragile and unstable media, including glass plates, nitrate and acetate-based film negatives, and magnetic audio and video tapes. The degrading analogue formats lead to unrecoverable information loss, can damage other archival materials, and in the case of nitrate negatives, may pose health risks and environmental hazards. Paul Conway emphasizes the fact that the preservation of audiovisual collections remains a major challenge of the twenty-first century and points out that the efforts of preservation community in preserving paper-based materials have not been extended to audiovisual resources.⁴ Digitization offers an opportunity to recover the content of deteriorating film negatives and tapes and to extend the usefulness of audiovisual materials as information sources.

This paper explores the issues of digitization and preservation of endangered historic photographic collections. The focus of the paper is on cellulose nitrate film negatives, a particularly challenging format because of its inherently unstable and highly flammable nature. The authors provide an overview of the debate on digitization as a preservation strategy and contribute to this discussion by describing the digitization project undertaken at the American Geographical Society (AGS) Library. The two-year project (2010 –2012), funded by the grant from the National Endowment for Humanities aimed at preserving and providing access to over 69,000 nitrate negatives from the historic photo collections at the AGS Library.

2. Digitization as a Preservation Strategy

Digitization as a form of preservation met with strong skepticism in the preservation community, especially in the early stages of digitization and digital library development. Digital conversion was viewed as an approach for creating surrogates for access and reproduction, but not as a preservation method. The benefits of digitization in expanding access and facilitating new type of use were widely acknowledged.⁵ Abby Smith in the 1999 publication *Why digitize?* emphasizes “digitization is access – lots of it,” but at the same time adds “digitization is not preservation, at least not yet.”⁶ Several years later, Janet Gertz expresses a similar opinion, but also notes that digitization can aid preservation by protecting fragile and valuable analogue materials from additional handling. She acknowledges that a digital copy

¹ Abby Smith, *Why Digitize?* (Washington, D.C.: Council on Library and Information Resources, 1999), 6; Steven Puglia, Jeffrey Reed, and Erin Rhodes, *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images* (College Park, MD, U.S. National Archives and Records Administration, 2004), 1.

² Laura Capell, “Digitization as a Preservation Method for Damaged Acetate Negatives: A Case Study,” *The American Archivist* 73 (2010): 246.; Paul Conway, “Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas,” *Library Quarterly* 80 (2010): 72-73.

³ *Ibid.*, p. 72.

⁴ *Ibid.*

⁵ Michael Lesk, *Understanding Digital Libraries* (Boston: Elsevier, 2004).; Smith, *Why Digitize?*

⁶ *Ibid.*

can serve as the only record of an original object that deteriorates or is destroyed, but still maintains that digitization is a form of copying, not preservation.⁷

The concerns about using digitization as an option for long-term preservation of analogue materials focus on the integrity and authenticity of digital data as well as the stability of digital formats and storage medium. In contrast to the established preservation methods, such as microfilming, paper facsimiles, or photo duplication, digital technology is relatively new and raises questions about access and retrieval of digitized data due to the possible obsolescence of hardware and software. Digital conversion projects undertaken according to digitization standards not only provide copies for immediate access and use, but also create a new valuable asset in the form of archival master files that require long-term digital preservation.⁸

The endorsement of digitization as a preservation reformatting strategy by the Association of Research Libraries (ARL) in 2004 represents a turning point in the discussion, although its focus is primarily on paper-based materials.⁹ The document recognizes digital conversion as one of the viable preservation options and points out that each preservation reformatting approach has its strengths and weaknesses. Microfilming, for example, a method often recommended because of the medium longevity, is at the same time limited in distribution, access, functionality, and characterized by low user adoption and satisfaction. The authors of the endorsement try to address the concerns of the preservation community to a certain extent by emphasizing the progress in standardization of file formats, the development of digitization standards and best practices, the commitment of the cultural heritage institutions to digital stewardship and the preservation of digital objects, and finally, a growing experience in refreshing and migration of digital data.

Digitization as a preservation method has been gradually gaining acceptance among the members of the cultural heritage community. The Preservation Reformatting Division of Library of Congress considers digital reformatting as a preservation method for at-risk archival materials among other options, such as microfilm and paper facsimiles copies.¹⁰ Deana Marcum emphasizes that the Library of Congress takes advantage of digitization to meet preservation needs and pays a particular attention to preservation of audiovisual materials.¹¹ The Endangered Archive Programme (EAP) at the British Library supports digitization as **preferred means of copying** of archival materials that are in danger of destruction or physical deterioration. This recommendation is particularly relevant in developing countries where other preservation methods, such as microfilming may not be available. In fact, digitization has increased the awareness of preservation and conservation issues and made possible to create copies of endangered archival and library materials worldwide. The EAP's guidelines emphasize the quality of digital images

⁷ Janet Gertz, "Preservation and Selection for Digitization," Northeast Document Conservation Center, 2007, accessed August 1, 2012, <http://www.nedcc.org/resources/leaflets/6Reformatting/06PreservationAndSelection.php>

⁸ *National Information Standards Organization, A Framework of Guidance for Building Good Digital Collections*, 3rd ed. (December, 2007), 25-35.

⁹ Kathleen Arthur, Sherry Byrne, Elisabeth Long, Carla Q. Montori, and Judith Nadler, "Recognizing Digitization as a Preservation Reformatting Method," *Microform and Imaging Review* 33, no. 4 (Fall 2004): 171-180. Prepared for the Association of Research Libraries (ARL) Preservation of Research Library Materials Committee.

¹⁰ Library of Congress Preservation Reformatting Division, "Preservation Digital Reformatting Program," accessed August 2, 2012, <http://www.loc.gov/preservation/about/prd/presdig/index.html>.

¹¹ Deanna B. Marcum, "Digitizing for Access and Preservation: Strategies for the Library of Congress," *First Monday* 12 (July 2, 2007), accessed August 1, 2012, <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/rt/prINTERfriendly/1924/1806>.

and standardized formats to facilitate long-term preservation of digitized objects.¹² Even the critics of the ARL endorsement recommend using digitization as a reformatting method where it offers the best or only chance for saving endangered information sources.¹³

The traditional concepts of saving and sharing valuable and fragile resources undergo a significant transformation in the digital world. Handling of rare items is less of an issue if there are multiple digital copies available for access. Archival digital master files provide high quality digital representations that can serve as preservation copies for a long time, especially if the concerns over the digital preservation continue being addressed. As Clifford Lynch points out, “digitization can provide a form of insurance for preserving content.”¹⁴ Sharing does not simply imply presenting an original artefact or its copy, but also involves providing access points and an extended description, creating a meaningful presentation and bringing together dispersed resources, and in some cases, enabling user participation and contribution. Those additional activities, although not traditionally associated with preservation, make digitized objects more valuable and useful research items.

The debate on digitization for preservation shifts from the emphasis on reformatting to the issues of usefulness and quality of preserved items, and to the crisis in preservation of audiovisual collections.¹⁵ Paul Conway examines the concept of preservation and makes a clear distinction between digitization for preservation and digital preservation. He defines digitization for preservation as “activities that result in the creation of digital products worthy of long-term preservation,” while digital preservation is understood as a set of policies and technologies aimed at protecting the value of objects created as a result of digitization as well as those that are born in the digital format.¹⁶ The activities associated with digitization for preservation include not only the conversion process, but also selection, creation of full and accurate description, and digital collection building. Digital assets worthy of digital preservation are comprised of high-quality digital master files and added intellectual value.

2.1 Digitization for Preservation of Film-Based Photographic Materials

The concept of usefulness is particularly relevant in the context of digitization for preservation of deteriorating film-based photographic collections. In the discussion of preservation in the “age of Google,” Conway brings up his earlier definition that points to the dual nature of preservation, actions that not only slow down the deterioration of archival materials, but also restore their usefulness as an information source.¹⁷ Audiovisual resources represent a major preservation challenge due to their inherently unstable physical nature but their usefulness is also limited because of the lack of access points and difficult-to-access formats. The focus of the following discussion is on film negatives of still photographs. The issues of preservation and intellectual control of other non-book materials, such as film-

¹² British Library, “The Endangered Archives Programme,” accessed August 1, 2012, <http://eap.bl.uk/pages/about.html>.

¹³ Andrew Hart, “A Critique of ‘Recognizing Digitization as a Preservation Reformatting Method’,” *Microform and Imaging Review* 33, no. 4 (Fall 2004): 187.

¹⁴ U.S. National Commission on Libraries and Information Science (NCLIS), “Mass Digitization: Implications for Information Policy” (Report from “Scholarship and Libraries in Transition: A Dialogue about the Impacts of Mass Digitization Projects” Symposium, March 10-11, 2006, Ann Arbor, MI), May 9, 2006, accessed August 3, 2012, <http://www.lib.umich.edu/mdp/symposium/NCLIS-report.pdf>.

¹⁵ Conway, “Preservation in the Age of Google,” pp. 64-74.

¹⁶ *Ibid.*, p. 65.

¹⁷ *Ibid.*

based moving image, audio, and video tapes are related and perhaps even more challenging, but are out of scope of this paper.

Prior to the development of digital photography, still photographs were recorded on a variety of analogue formats including glass plates and different types of film from nitrate to acetate and polyester negatives. Photographers would print a limited number of images for publications and exhibits, but the vast majority of photographic images remained on a difficult-to-access film negatives. The inaccessibility of photographic film formats seems to be one of the major factors of limited awareness of those collections and their scholarly use.

In addition, archival image collections tend to be poorly organized, have limited item-level description, and generally lack individual indexing records and contextual information.¹⁸ Professional photographers and even scholars who took images during their scientific expeditions often did not document their collections very well. Very few photographic collections have annotations of individual images and even so, the level of consistency and accuracy of description vary from item to item. In library and archives settings, photographic resources are processed and described only on the collection level. James M. Turner contributes the lack of item-level indexing records to the perception of audiovisual resources as “second class materials.”¹⁹ Margery S. Long and Mary Lynn Ritzenthaler note a similar attitude among archivists and historians, who “did not always recognize photographs as primary source materials.” The authors add, “in the formative years of archives, only written records were regarded as archival and deserving preservation.”²⁰

Digitization draws attention to photographic collections that have been neglected over the years and offers an opportunity to address both aspects of their preservation. It allows us to capture and to preserve the content of deteriorating analogue formats and to create a new potential for their use. Creating descriptive metadata on item level for digitized objects and their integration into digital collections removes barriers to access and enables discovery of those unique visual resources in the digital environment. Digitization contributes to making “visible” a large body of historical visual evidence, as many of the images become available for public viewing for the first time.

Digitization as a preservation strategy, for example, has been successfully tried for a film-based photographic collection by the archivists at the University of Southern Mississippi.²¹ The goal of the project was to recover the visual content of highly damaged acetate-based film negatives from the Robert C. Waller Photograph Collection at McCain Library and Archives . After reviewing various strategies for capturing and preserving the visual content of the deteriorating negatives, the archivists determined that digitization was the most suitable option, despite the concerns over the long-term digital preservation.

Archival film-based photographic collections may include both acetate and nitrate negatives. Acetate negatives, part of the so-called safety film, were introduced in the mid-1920s with the intention of replacing the more hazardous nitrate film. The American Geographical Society Library includes in its holdings extensive photographic collections of both nitrate and acetate film negatives. Nitrate negative collections, however, were identified as a top preservation priority because of health and environmental hazards

¹⁸ Margery S. Long and Mary Lynn Ritzenthaler, “Photographs In Archival Collections,” in *Photographs: Archival Care and Management*, ed. Mary Lynn Ritzenthaler and Diane L. Vogt-O'Connor (Chicago: Society of American Archivists, 2006), 2; James M. Turner, “From ABC to http: The Effervescent Evolution of Indexing for Audiovisual Materials,” *Cataloging & Classification Quarterly* 48 (2010): 84-86.

¹⁹ Turner, “From ABC to http,” p. 86.

²⁰ Long and Ritzenthaler, “Photographs In Archival Collections,” p .2.

²¹ Capell, “Digitization as a Preservation Method,” pp. 235-236.

associated with nitrate film. The first large-scale digitization for preservation initiative at the American Geographical Society Library was undertaken to address the problem of deteriorating nitrate negatives.

3. The Problem of Cellulose Nitrate Negatives

The archives and libraries include collections of still photographs recorded on a variety of analogue formats. Nitrate film negatives represent a significant portion of archival holdings since this format was used by professional and amateur photographers for over 50 years. The introduction of cellulose nitrate negative film by Eastman Kodak in 1887 represented a revolution in photography.²² Nitrate negatives were more flexible and durable than previously used glass plates and allowed photographers to take more pictures and in a variety of conditions. Cellulose nitrate film was first used for roll films and later on for sheet films in a variety of formats. Nitrate film also presented fewer problems in breaking and was portable, which meant it could have been used for travel photography and for documentation of geographic and scientific expeditions. The advantages of nitrate film outweighed its drawbacks and contributed to its widespread use. Because of its durability, nitrate film negatives continued to be used by photographers long after the introduction of acetate based safety film. The production of nitrate sheet film ceased in the late 1930s. Eastman Kodak stopped manufacturing nitrate roll film in 1950.

Instability of cellulose nitrate film and safety risks were the main reasons for discontinuation of the production of nitrate negatives, but safety and preservation challenges have remained for vast photographic collections in the library and archives settings. Nitrate is highly flammable and the gas from burning nitrate is hazardous. Cellulose nitrate if ignited, cannot be extinguished. The film burns in the absence of oxygen producing its own supply.²³ The risk of fire cannot be underestimated as several fires did take place, especially with nitrate-based moving picture film, earning the nitrate film a reputation “the film stock from hell.”²⁴ There is also a relationship between nitrate film decomposition and combustibility as fires can be caused by spontaneous combustion of nitrate film in the late stages of decay.

All nitrate-based film inevitably decomposes with age leading to image alteration, eventual loss of visual content, and additional health and safety hazards. The preservation community has identified six levels of nitrate decomposition.²⁵ Nitrate negatives begin losing photographic detail at level three; at level six, the film turns into brownish acid powder. The rate of deterioration depends on environmental conditions and accelerates with the exposure to heat and humidity. During the decomposition process nitrate film produces acidic gases that can have a damaging effect on other library materials in close proximity, causing *embrittlement* of other film and paper. Nitric acidic gases can also create a potential health risk for library staff as extended exposure can cause rashes, nausea, headache, respiratory problems, and eye irritation.

The guidelines for care and storage in archival settings recommend separating cellulose nitrate negatives from other negatives and archival materials.²⁵ Since the decomposition of nitrate film varies greatly by manufacturer and production date, some nitrate film has survived in very good condition and

²² Monique Fischer, “A Short Guide to Film Base Photographic Materials: Identification, Care, and Duplication,” Northeast Document Conservation Center, 2008, accessed August 2, 2012, <http://www.nedcc.org/resources/leaflets/5Photographs/01ShortGuide.php>.

²³ Heather Heckman, “Burn After Viewing, or, Fire in the Vaults: Nitrate Decomposition and Combustibility,” *The American Archivist* 73 (2010): 483.

²⁴ *Ibid.*, p. 490.

²⁵ Fischer, “A Short Guide to Film Base Photographic Materials.”

some of it has degraded beyond use. Fluctuations and high levels of temperature and humidity also speed up the decomposition process thus, negatives should be kept in cold storage to slow down the deterioration. Photo duplication to stable safety film has been recommended as a traditional reformatting method for film-based photographic materials at risk of deterioration.²⁶ However, photo duplication is not feasible in large-scale conversion projects and does not address the issue of usefulness in regard to access points and description. Duplicating negatives onto safety film still keeps the images “hidden” from ready use. It is also not a suitable reformatting method for negatives in various stages of decomposition as photographic processes may not be able to capture the content from a damaged image.²⁷

After reviewing the available reformatting methods and concerns, the staff at the American Geographical Society Library decided to use digitization as an option for saving and sharing the large body of nitrate negative photographs in its holdings.

4. The Nitrate Negative Photographic Collections at the AGS Library

The American Geographical Society (AGS) Library has served as a repository of photographs taken by the AGS fellows and associated researchers during their expeditions around the world. The AGS Library, one of America’s oldest, largest and most distinguished geographical research libraries, was established in New York City in 1851. The Library and its photographic collections were moved to its new location at the University of Wisconsin-Milwaukee in 1978.²⁸ The AGS Archives with additional photographic resources was transferred to Milwaukee in 2011.

The Library’s photographic collections consist of 440,000 images in a variety of formats, including glass plate negatives, film negatives, slides, lantern slides, and prints. The collections include the work of historic photographers as well as explorers, surveyors, journalists, geographers, travelers, and scientists in various disciplines. The extensive photography collection of the AGSL dates from early photography in the 1860s to the present. Those historic photographs serve as a visual memory of the world as they document the history and culture of regions and countries where visual representation could have not been recorded or has been destroyed. The early photographs recorded on nitrate negatives are particularly valuable since they come from an era where camera ownership was not widespread and imagery is today scarce. In some cases landmarks have been destroyed so these images are even more significant as historical documents.

The nitrate negative collections of still photographs represent a significant portion of the AGS Library’s collections. Initially, it was estimated that AGS Library’s photographic collections contained approximately 52,000 nitrate negatives. However, with the identification and inventory undertaken as part of the digitization project the number of known nitrate negatives grew significantly, totaling 69,625. Some collections were found to include more nitrate negatives than originally thought. The Pendleton collection, for example, first estimated at 4000 images, yielded 16,817 by project end. In addition, other nitrate negatives and collections were discovered in the AGS-New York archives, of which the AGS Library became the custodian in 2011.

²⁶ Ibid.

²⁷ Capell, “Digitization as a Preservation Method,” p. 241.

²⁸ For more information about the AGS Library and its holdings, visit the library website at: <https://www4.uwm.edu/libraries/AGSL/>.

Photographic images were, for the most part, acquired to serve the research and publication needs of the American Geographical Society. Included were images from notable geographer/photographers such as Isaiah Bowman, Frederick Clapp, Helmut De Terra, Alexander Forbes, Richard U. Light, and Robert L. Pendleton. Many of the AGSL's nitrate negatives are from these sources. Other collections, such as those of Theodore deBooy, William O. Field, Mary Jo Read, and Harrison Forman were donated to the AGS Library by the photographers or their families. The nitrate negative holdings contain a large number of unique images that have never been published and rare images that cannot be easily accessed.

The AGS nitrate negative collections represent almost 40 years of documentary photography starting with Isaiah Bowman's photographic record of Yale South American Expedition in 1907 to Bert Krawczyk's photographs in the Yunnan Province in China between 1943 and 1945. The images span all continents, with the exception of Antarctica, and document a global range of peoples, cultures, and landscapes as seen through the eyes of geographers and photojournalists. The highlights of the collections include a remarkable set of early twentieth century images of the archaeological sites in Tibet, India and Chinese Turkestan from the Helmut De Terra Collection; images of Iran and Afghanistan by Frederick Clapp in the 1930s; William Field's photographs of Russia and the Caucasus Region from 1929-1933; the first aerial photographs of Africa and South America taken by Mary Upjohn Light Meader (1937); Robert L. Pendleton's extensive photography of Thailand from 1930s-1940s; and a unique photographic record of Tibet and China in the 1930s provided by a photo journalist, Harrison Forman.

Nitrate negatives were identified in 17 different photographic collections held at the AGS Library. The numbers of nitrate negatives vary from collection to collection with several hundred in Mary Jo Read Collection, for example, to over 15, 000 in the Harrison Forman or over 16, 000 in the Pendleton Collection. Some collections contain exclusively nitrate negatives, while in others nitrate film was interfiled with safety film and required additional identification, separation, and processing. Prior to undertaking the digitization project, approximately two thirds of the nitrate negatives in the AGS Library's holdings were stored in original envelopes and containers. Large collections of nitrate negatives were separated from other library materials and stored in an isolated room, but none was placed in cold storage.²⁹ The digitization project provided an opportunity to capture the visual content of deteriorating negatives and at the same time to address the preservation and conservation of the original source materials.

5. The Digitization Project: Saving and Sharing the Nitrate Negative Images

The two-year project (2010 –2012), "Saving and Sharing the AGS Library's Historic Nitrate Negative Images" was funded by a grant from the National Endowment for Humanities. The goals were to reformat, provide access, safely re-house the AGS Library's large collection of cellulose nitrate photographic negatives and ensure their proper storage, and provide long-term preservation of their digital representations. The project built on the library staff's previous experience in digitization, the knowledge gained from the prior preservation survey of AGS Library's photographic collections, and a successful pilot project. The pilot digitization project was an excellent primer for setting up a system to safely handle, digitize, and preserve the negatives for long-term storage.

²⁹ Fischer, "A Short Guide to Film Base Photographic Materials," p. 4.

The focus of the two-year project was the digitization for preservation of all nitrate negatives in the AGS Library's holdings. The project involved a number of activities that aimed at preservation and providing access, including:

- Identification and inventory of nitrate negatives in the AGS Library's photographic collections
- Preparation of nitrate negatives for scanning and re-housing
- Digital reformatting with creation of archival master files for long-term preservation and derivative images for access
- Creating item-level metadata for all digitized objects
- Integrating selected images with associated metadata into digital collections
- Packaging and placing nitrate negatives in cold storage
- Transferring archival master files to the campus digital repository for long-term preservation
- Documenting the project

5.1 Preparation and Re-housing of Negatives

Most of the nitrate negatives in AGS Library's collections identified for this project were stored in the original envelopes and containers and required processing and re-housing prior to undertaking digital reformatting. Approximately a third of the collections were processed according to archival standards prior to undertaking digitization or as part of the pilot digitization project. The majority of negatives, however, were either enclosed in brittle, deteriorating envelopes, or rolled in cans or over acidic paper cores. Please see Figure 1 for an example of a collection of nitrate negatives stored in a metal box and brittle paper enclosures.



Figure 1. Nitrate negatives stored in original paper envelopes.

The process of re-housing consisted of inserting the negatives into acid-free envelopes, assigning identifiers, and placing the envelopes into archival boxes. As demonstrated in Figure 2, each negative was sleeved in buffered, acid free, lignon free, paper enclosures. An original id, if available and a unique digital id were recorded on the envelope. The envelopes were in turn placed into acid free storage boxes. When necessary, negatives were cleaned before re-housing and before scanning. Long strips of film were unrolled, relaxed, and cut into sections to fit standard sized envelopes.



Figure 2. Re-housing of nitrate negatives.

Nitrate film in the Platt and Forman collections posed major challenges during the preparation and re-housing process. Roll film in these collections was uncut and rolled around oblong cardboard cores or rolled into metal cans. This film had been rolled onto cores or in cans for over 60 years and it was resistant to straightening by traditional means without damaging it. The project staff found that mild humidification would relax the film sufficiently to unwind it and then roll it in reverse.

5.2 Digital Reformatting

The goal of digital reformatting was to capture the visual content of the nitrate negatives and to create high-quality digital master files, “worthy of long-term preservation.”³⁰ The guidelines for the project were based on digital library standards and best practices to ensure a consistent level of image quality and to create digital objects in support of current and future use.³¹ Digital master files were created as a direct result of the image capture process. General recommendations for digital master file creation included:

- Scanning at the highest quality affordable
- Assigning a unique Digital ID
- Using TIFF as a non-proprietary format
- Saving the original scan without any enhancements
- Saving digitized images with no compression

The scanning resolution was based on the size of the original negative. For example, 35mm film was scanned at 4000 pixels per inch (ppi) on the long side, while sheets of 8x10 in. were scanned at 400 ppi. All individual scans were assigned a unique file name (Digital ID) based on the previously established

³⁰ Conway, “Preservation in the Age of Google,” p. 65.

³¹ BCR’s CDP Digital Imaging Best Practices Working Group, “BCR’s CDP Digital Imaging Best Practices,” Version 2.0 (June 2008), p. 7; “A Framework of Guidance for Building Good Digital Collections,” p. 26.

AGS Library's file naming convention. Digital IDs were also inscribed on the envelopes to provide reference between a negative and its digital representation.

The original project plan was to outsource scanning of all negatives to a digitization vendor. However, this plan had to be modified because of the challenges in preparing, packing, and transportation of nitrate film. Nitrate negatives are considered a hazardous material, and as such, transporting them imposed rigorous packaging standards that took extra time and required purchase of expensive packing containers. In addition, processing of negatives in a variety of sizes proved to be very time consuming. In the case of the Forman Collection the photographer's sequencing intermingled the various photographic formats in which he worked. To have sent these images to the vendor would have required separating the formats, which would have led to disarrangement and mismatch of picture and description. To obviate this risk, the Libraries purchased a new Hasselblad scanner and digitized this collection in-house. All in all, of the 69,625 negatives scanned, 50,655 were digitized by the vendor and 18,970 were scanned in-house.

Digital reformatting provided an opportunity to recover the visual content of the negatives. The scanning process also demonstrated the problems with nitrate film decomposition and revealed the degree of image degradation. Most of the nitrate negatives in the AGS Library's holdings were in a relatively stable condition, suitable for scanning. In fact, curly roll film presented more challenges for digital conversion than decomposed negatives. There were only few instances of advanced level of film decomposition where the negatives were too disintegrated to be scanned.

The loss of visual content due to nitrate film instability, however, became evident in many images after scanning. A significant portion of the digitized images exposed some degree of fading and even the loss of legible photographic detail. The rate of deterioration varied from collection to collection and even from negative to negative within the same collection. The extensive Harrison Forman Collection proved to be in relatively stable condition, although Forman used nitrate film in his photojournalistic work for over a decade and tried a variety of formats including rolls and sheet negatives. Digital conversion of his collection rendered a large body of high-quality images with a minimum loss of photographic detail. On the other hand, 40% of the images scanned from the Mary Jo Read Collection turned out to be illegible or revealed significant image alteration (see Figure 3 for an example of a faded image). The information about image degradation was recorded in the project database. The images that displayed a considerable loss of visual content were not included in digital collections for public access.



Figure 3. Image scanned from a nitrate negative showing signs of film decomposition.

All digital master files created as a result of digital reformatting have been retained regardless of the image deterioration detected during the conversion process. All scans were reviewed for quality and saved as uncompressed TIFF files. They serve as preservation copies and a source for derivative images. A second copy of TIFF files, so called “service file” was created if digitized images had to be processed to remove dust marks, scratches, and to improve contrast of faded images. The archival master files remain uncompressed, while the service copy is preserved as LZW lossless compressed TIFF. Both digital master files and service copies were transferred to the campus digital repository for long-term storage. The archival master files and service copies are stored in the University of Wisconsin-Milwaukee digital repository and follow standard campus procedure for archiving, backup, and migration. University of Wisconsin-Milwaukee Libraries that houses the AGS Library is committed to long-term preservation of digital master files, long-term archival preservation of original nitrate negatives, and sustaining the digital collections created as result of this project.

All digital master files created as a result of digital reformatting have been retained regardless of the image deterioration detected during the conversion process. All scans were reviewed for quality and saved as uncompressed TIFF files. They serve as preservation copies and a source for derivative images. A second copy of TIFF files, so called “service file” was created if digitized images had to be processed to remove dust marks, scratches, and to improve contrast of faded images. The archival master files remain uncompressed, while the service copy is preserved as LZW lossless compressed TIFF. Both digital master files and service copies were transferred to the campus digital repository for long-term storage. The archival master files and service copies are stored in the University of Wisconsin-Milwaukee digital repository and follow standard campus procedure for archiving, backup, and migration. University of Wisconsin-Milwaukee Libraries that houses the AGS Library is committed to long-term preservation of digital master files, long-term archival preservation of original nitrate negatives, and sustaining the digital collections created as result of this project.

5.3 Cold Storage

Although all of the film has been scanned at a high resolution, it was decided at the outset of the project that the original nitrate negatives would be preserved for at least the foreseeable future. Re-housing was the necessary first step in the process, but long-term archival preservation of the negatives would require maintaining the film at low temperatures. Sleeved, labeled, and boxed negatives were packaged according to the latest research on the cold storage of nitrate negatives.³² Figure 4 provides an example of boxes of nitrate negatives prepared for cold storage. The boxes are stored in two cold storage units installed in the library, including the explosion-proof freezer, which is being used for nitrate negative collections that are in later stages of degradation.

5.4 Creating Descriptive Metadata on Item Level

Metadata creation represented a significant part of the digitization project and was undertaken to provide access points to digitized images and to extend their usefulness as information resources. Metadata records were created on two levels:

³² Sarah S. Wagner, “Cold Storage Options: Costs and Implementation Issues,” *Topics in Photographic Preservation* 12 (2007): 1-10.



Figure 4. Boxes of nitrate negatives prepared for cold storage.

- Minimum item-level metadata were recorded for all digitized negatives
- Extensive descriptive metadata were created for images selected for publishing in the AGS Library's digital collections

The minimum metadata were recorded for all 69,625 digitized negatives to capture photographers' original designation and descriptive notes. The data was entered into the locally stored database. In addition to the photographer's note, the minimum records include date of creation, original negative ID or accession number, Digital ID, notes on the negative condition, nitrate negative location, and an original collection name. Digital IDs serve as record identifiers since the numbers or descriptors assigned to original negatives were not unique. The minimum metadata records formed the basis of more extensive descriptive metadata for selected images.

The images selected for online publication in the AGS Library's digital collections were provided with extensive metadata. The intent of metadata was to record and present all pertinent descriptive and administrative information regarding a particular scanned image including: what it is, its location in the world, who is responsible, its physical attributes, where it came from, where it is now, when the image was created, the intellectual property rights, its digitization details, grant funding credit and where it now resides online and as a physical object.

The completeness of the descriptive metadata varies from collection to collection reflecting the detail of the photographer's annotations. In some cases, negatives came with little or no descriptive information but were rich historical images that warranted additional research (see Figure 5 for an example of limited original item description). The metadata records of some collections were augmented with ancillary information. In the case of Forman and Field Collections, for example, the photographers' field diaries served as an additional source of descriptive information. The diaries were digitized as well and are presented as part of the AGS Library's digital collections as valuable adjuncts to the photographic images.

The metadata structures for AGS Library's digital collections are based on a qualified Dublin Core schema. The natural language field labels in the public display are internally mapped to Dublin Core schema to insure cross collection searching and metadata harvesting. The customized metadata template consists of 38 fields, including descriptive elements, such as Title, Date of Photograph, Photographer's Note, Photographer Name, Description, Related Resources, and two subject fields. The two fields are designated to capture different concepts and use different controlled vocabulary tools: Subject TGM covers topical subjects and derives terms from the Library of Congress Thesaurus for Graphic Materials



Figure 5. Limited description of original negatives in the Harrison Forman Collection.

(LC TGM) while Subject LC indicates proper names of people and objects depicted in images and uses Library of Congress Subject Headings (LCSH).

The description of geographic coverage required special attention in creating metadata records because of the nature of AGS Library's photographic collections documenting geographic expeditions and scientific exploration around the world. The fields related to geographic location include Continent, General Region, Country, Region, Province/State, County/Municipality, City/Place, and Geographic Feature. The controlled vocabulary is selected from the Getty Thesaurus of Geographic Names (TGN). The terms are not pre-coordinated and the records are not always complete due to the gaps in the original item description. The metadata creators often needed to conduct in-depth research on the geographic details because the need for geographic specificity frequently exceeded the extent of published thesauri.

In addition, metadata records include information about the size, medium, location, and provenance of nitrate negatives creating a link between original items and their digital representation. As demonstrated in Figure 6, metadata records created as a result of this project provide rich descriptive information for selected images and enable access to a large body of historical visual record that was previously inaccessible.

5.5 Digital Collection Building

Digital collection building involved selecting images, extracting metadata records from the local database, and uploading the images with associated metadata into CONTENTdm, the digital collection management system used for constructing AGS Library's digital collections. Derivative images in the JPEG2000 format were created during this process. Approximately 64% of the digitized images were selected for the online publication in the AGS Library's digital collections.³³ The reasons for elimination were: 1) poor image quality that was the result of nitrate film decomposition or the initial defect, such as poor focus or light balance, etc.; 2) duplication that is related to the fact that photographers would often shoot several similar pictures; 3) inappropriate subjects: some of our photographers had scientific

³³ The average selection rate does not include 23,780 images from the Robert S. Platt Collection. These images were scanned for preservation as part of the grant project, however metadata creation has been included in the AGSL's subsequent grant request: "Saving and Sharing the AGS Library's Historic Film Collections II: Monochrome Acetate Negatives and Motion Picture Film."



Title	Shanghai (China), aftermath of Palace and Cathay hotels bombing
Part of Set	Harrison Forman Collection - China
Date of Photograph	1937-08-14
Photographer's Note	China: Shanghai. Sent via Pres. Jefferson (ship) - the 1st Dollar Line steamer evacuating American worn leaving Manila on August 20th. Pix of bombing of CATHAY and PALACE HOTELS, Shanghai, on August 13
Photographer	Forman, Harrison, 1904-1978
Subject TGM	People Dead persons War casualties War damage Bombardment
Subject LC	Sino-Japanese War, 1937-1945--China--Shanghai
Continent	Asia
General Region	East Asia
Country	China
County/Municipality	Shanghai Shi (municipality)
City/Place	Shanghai
Type	Image
Original Collection	Harrison Forman Collection
Original Item Size	35mm
Original Item Medium	Black-and-white negatives cellulose nitrate film
Original Item Location	Box 20
Provenance	Gift of Sandra Carlyle Forman
Repository	American Geographical Society Library, University of Wisconsin-Milwaukee Libraries
Rights	The Board of Regents of the University of Wisconsin System
Digital Publisher	University of Wisconsin-Milwaukee Libraries
Digital ID	fr208082
Digital Collection	American Geographical Society Library Digital Photo Archive: Asia and Middle East
Digital Format	Image/JPEG
Project Name	NEH Grant Project: Saving and Sharing the AGS Library's Historic Nitrate Negative Images
Part of	American Geographical Society Library Digital Photo Archive

Figure 6. An example of a digitized image from the Harrison Forman Collection with associated metadata.

specialties such as petroleum geology or soil science which were heavily represented in their photographic output. Such photos, while of potential interest to the specialist, were deemed inappropriate for extensive publication online, especially in a humanities-focused website.

The digital objects created as a result of this project were added to several digital collections based on their geographic location. Digital collections of the AGS Library are arranged by geographic location according to continents following the standards of geographic organizations. This approach resulted in six major portals, including Africa, Asia and Middle East, Europe, North and Central America, South

America, and Oceania. The decision to organize the collections by geography rather than by a photographer or collection's name was undertaken to meet the information needs of end-users who tend to search by subject or geographic location. This approach also allows the AGS Library to bring together images of the same location from different time periods and from multiple collections. The AGS Library's digital collections are open to the public and are available at: www4.uwm.edu/Library/digilib/index.html. In addition, the website dedicated to the project, *The NEH Grant to Preserve Nitrate Negatives in the AGSL* provides descriptions of individual collections and access to records by photographers' names and major topical themes. Individual metadata records also include links to original collections and photographers' names. The portal is available at: www.uwm.edu/libraries/digilib/NEHgrant/.

6. Conclusion

This paper contributes to the debate on digitization for preservation by exploring the ways in which digitization brings a renewed attention to preservation of endangered historic photographic collections and enables representing and preserving the content of fragile and deteriorating visual materials. It supports the view that preservation not only protects deteriorating archival materials but also restores their usefulness as information resources. The concept of digitization for preservation is discussed in light of a large-scale digitization project aimed at preserving and providing access to over 69,000 nitrate negatives at the American Geographical Society Library. Digitization as a preservation strategy can only be considered as a framework that ensures the creation of digital master files according to the digitization standards, extends the usefulness of digitized objects through providing full and accurate description, and addresses sustainability and long-term preservation of digital masters.

Born Digital Images

Creation to Preservation

Jessica Bushey

School of Library, Archival and Information Studies (SLAIS) The iSchool at University of British Columbia

Abstract

In the last decade professional photographers have adopted born digital photographic processes to fulfill business needs and realize creative endeavours. More recently, the proliferation of born digital images on social media websites as documentation of individual experiences and a primary tool for communication has highlighted the necessity for information professionals to understand the key factors affecting the reliability and authenticity of born digital images as records. Based on findings from the International Research on Permanent Authentic Records in Electronic Systems (InterPARES 2) “Survey on the Recordkeeping Practices of Photographers using Digital Technology”; this paper discusses the key factors affecting the reliability and authenticity of born digital images such as image file formats, metadata, workflow procedures, and storage media. In an effort to support cultural institutions responsible for the acquisition, management and preservation of born digital images, this paper provides an overview of recent developments initiated by photographers, the imaging industry and cultural heritage organizations. These developments reflect the dynamic state of digital practice and reveal the importance of collaboration among photographers, software developers, hardware manufacturers, and cultural stakeholders in the creation and preservation of born digital images.

Author

Jessica Bushey is a doctoral student at the School of Library, Archival and Information Studies (SLAIS) The iSchool at University of British Columbia (UBC). Her research interests address the trustworthiness of born digital media and authorship in the online environment. Ms. Bushey is a Systems Analyst for Artefactual Systems Inc., the lead software developer for open-source archival applications ICA-AtoM and Archivematica. After receiving her MAS in 2005, Ms. Bushey developed and implemented the digitization program at the Museum of Anthropology at UBC. This four-year project involved born digital capture of 35,000 artefacts in order to increase public access to the collection, encourage scholarly research, and support preservation activities.

“It took 50 years for us to reach 14 million analog images in the Still Picture Unit, but we’ll reach 10 million digital images in approximately 10 years.”

*-- Billy Wade, Archivist
Still Pictures Unit, U.S. National Archives¹*

1. Introduction

In the last decade we have witnessed the transformation of photography from a film-based medium to pixel-based. The switch to digital began with professional photographers working in fields such as

¹ Billy Wade, “From Analog to Digital,” *Media Matters: The Blog of the National Archives’ Special Media Archives Services Division*, April 18, 2012, <http://blogs.archives.gov/mediamatters/2012/04/18/from-analog-to-digital/>.

medicine, journalism, geospatial, and law enforcement. Supported by new technologies for capturing, transmitting and storing born-digital images, the practice of digital photography grew quickly and soon spread beyond professional boundaries and throughout the general populace. In fact, recent advances in cell phone/camera technology have managed to couple quality with convenience and as a result, unprecedented numbers of born-digital images are being created and shared on a global scale.

Unlike their analogue predecessors, digital images require software and hardware for viewing, sharing and storing. The convenience of traditional photofinishing services that added numbers to slide frames, provided protective housing for film and prints, and offered organizational materials at point-of-purchase are no-longer available. Contemporary photographers working with digital media must be self-reliant and perform the majority of management tasks with commercially available software (often a suite of applications) to the best of their abilities. Passive approaches to digital image management and preservation place valuable image collections at risk. With the exception of nitrate film, the storage requirements of early photographic materials provided generous margins for neglect; whereas, unnamed and forgotten born-digital images are unlikely to survive beyond the next upgrade. In light of emerging responsibilities for photographers to manage digital images and provide ongoing access to digital files it is necessary to re-articulate the role of the photographer as both author/creator and preserver.

This paper discusses the findings of the *InterPARES 2* general study entitled “Survey on the Record-Keeping Practices of Photographers using Digital Technology” in the context of industry initiatives aimed at the creation, management, and preservation of born-digital images.² An analysis of the *InterPARES 2* findings will establish the framework in which photographers are approaching digital practice and provide the foundation for an exploratory discussion of recent studies and current practices within the archival community aimed at preserving born-digital cultural heritage. Of particular importance is an examination of metadata schema standards used by professional photographers and supported by software vendors, and storage providers (including cloud services), to determine their efficacy to ensure born-digital images as reliable and authentic records.

2. Terminology³

The term “born-digital image” refers to a digital image that never physically existed before becoming a digital file. The most common example is an image created with a digital camera. An existing photograph

² InterPARES 2 Project, “General Studies,” accessed on August 31, 2012, http://www.interpares.org/ip2/ip2_general_studies.cfm. A general overview and the final report of the GS-07 Survey on Recordkeeping Practices of Photographers Using Digital Technology, can be downloaded from the project site, see http://www.interpares.org/ip2/ip2_general_studies.cfm?study=29. The IP2 GS-07 Survey was undertaken by Marta Braun (Ryerson University) and research assistant Jessica Bushey (University of British Columbia) and conducted under the auspices of the InterPARES 2 Project. For a more in-depth analysis of the findings of the “Survey on the Record-Keeping Practices of Photographers Using Digital Technology” see Jessica Bushey, “Born Digital Images as Reliable and Authentic Records,” MAS Thesis, University of British Columbia, Vancouver, 2005. An earlier version of the findings of the Survey was presented in the article, “He Shoots, He Stores: New Photographic Practice in the Digital Age,” *Archivaria* 65 (Spring 2008): 125-149. For the purposes of the current paper, the discussion has been significantly revised and updated to reflect changes introduced by both the photographic and archival communities in the past five years and to address the context of online networks including social media sites.

³ Parts of this section originally appeared in Bushey, “He Shoots, He Stores,” pp. 131-132.

or document that is scanned or digitally reproduced to create an image file is not considered to be born digital but to have been digitized.⁴

For the purposes of this paper, reliability is viewed as the trustworthiness of an image as a statement of fact and refers to the accuracy of its content. The accuracy of content is determined by the methods employed in the creation of the image. Examination of the controls over the procedure of creation and the authority and competency of the persons involved in these activities determines the reliability of an image. Authenticity refers to the fact that an image is what it purports to be and has not been tampered with or corrupted since it was set-aside. To ensure authenticity the integrity and identity of a digital image must be established and maintained. Establishing the integrity and identity of a digital image requires specific contextual information to remain linked to and/or embedded in the image file.

Lastly, the relationship between a born-digital image and a record must be addressed. From its nascence, photography has been associated with the act of recording an event. In the context of this approach, the photograph is received as a visual account of something and an aid to memory. The process of naming, saving, and setting-aside a digital image file for long-term storage makes explicit the creator's intent to carry forward visual information about an event for future use and/or reference. Additional capture of technical and descriptive metadata about the born-digital image further supports its capacity to function as a record.⁵

3. Survey on the Record-Keeping Practices of Photographers Using Digital Technology

Under the auspices of InterPARES2, the "Survey on the Record-Keeping Practices of Photographers Using Digital Technology" was launched as a web-based questionnaire in the Fall 2004. The survey targeted photographers who were known to create digital images and use digital technology to manage and store their images. An invitation to participate was posted to professional online fora and photographic association web sites that foster a community of photographers using digital technology.

The survey was contextualized within the larger research goals of InterPARES 2, mainly the investigation of problems surrounding the reliability, authenticity, permanence, and accessibility of digital records. The survey questions were formulated to gather information regarding the principles and procedures that contribute to the creation, use, and preservation of digital images as reliable and authentic records; however, the terms "reliability" and "authenticity" were omitted from the survey to avoid confusion resulting from individual and disciplinary interpretation of their meaning. Data analysis was conducted on the basis of qualitative techniques, such as tallying the responses to each multiple-choice question and expressing these numbers in percentages, and examining the additional textual responses for categories and themes.

⁴ The digitized image acts as a surrogate to the analogue original. A born-digital image refers to an image that is generated entirely from digital hardware and/or software, and which has no analogue genesis.

⁵ A record is a residue of activity retained by its creator for reference or use in later activity. See Luciana Duranti and Kenneth Thibodeau, "The Concept of Record in Interactive, Experiential and Dynamic Environments: the View of InterPARES," *Archival Science* 6 (2006): 13-68, p. 66.

4. Research Questions

The survey addresses the following questions: (1) What kinds of digital images do photographers produce? Of these digital images, which constitutes the “original?” (2) What are the assumptions of photographers about future access to their images? Additionally, what is the intention of photographers for the dissemination and presentation of their digital images? (3) What is the nature and variety of digital materials used by photographers? Specifically, what hardware and software do photographers use, and what methods or materials do they select for long-term storage?

5. Survey Findings⁶

Results of the survey are presented in two broad areas that reflect the method in which photographers approach digital image creation, use, and preservation. The first section addresses the actions and procedures that photographers use to create digital images, such as the selection of capture hardware and software, image file formats and their characteristics, and the automatic and manual addition of technical and administrative metadata. The choices made by photographers at this stage of their digital practice affect the reliability of the born-digital image. The second section addresses the steps taken by photographers to store and preserve their digital images, such as security measures, procedures for transmission and dissemination, selection of storage media, and the manual addition of administrative and preservation metadata. The choices made by photographers at this stage of their digital practice affect the authenticity of the born-digital image.

Throughout both sections the critical role of metadata, and its essential contribution to establishing the reliability and proving the authenticity of a born-digital image, will be discussed because metadata are automatically generated and manually input throughout the life cycle of a record. This paper is concerned with metadata schemas that are accepted standards and currently available to photographers via the functionality of hardware and software products. Where applicable, developments instigated by the photographic community will be discussed, including changes in practice and adoption of new metadata schemas.

6. Creation and Use

The majority of survey respondents identified their practice as “completely digital” and provided additional comments that date their transition to digital as commencing in 1998. The ability to re-purpose digital images (i.e., re-format and share the same image to serve different creative and business needs) influences most photographers’ choice of image file formats at the time of digital capture. Many professional photographers select proprietary formats, such as *RAW*, for initial capture because it offers the highest quality data with the most potential for re-purposing. The *RAW* format is exclusive to *Digital Single Lens Reflex (DSLR)* cameras marketed to professional photographers. In cases where photographers use *RAW* format to capture the scene digitally, the *RAW* file is equated with the ‘original’ image and treated as such throughout subsequent procedures for use and preservation. The fact that *RAW* image formats are proprietary, often differing for each successive camera model from the same

⁶ Parts of this section originally appeared in Bushey, “He Shoots, He Stores,” pp. 133-141.

manufacturer, and that the practice of encryption to conceal the RAW specification from users is widespread, present an enormous risk to the future usability of born-digital images created and stored in the RAW format. In 2004 Adobe Systems Inc., launched the *Digital Negative (DNG)* format as an alternative RAW format aimed at supporting image preservation. The DNG is based on the *Tagged Image File Format (TIFF)* specification and provides photographers a file format that is self-contained (image and metadata intact) and cross-platform interoperable.⁷ In 2005 the OpenRAW initiative was launched by photographers to raise awareness of the risks posed by proprietary RAW formats to digital image preservation. In 2006 OpenRAW conducted a survey of over 19,000 photographers to gather data about professional practices and concerns regarding RAW formats. One of their key findings is the “increased probability that as time passes a RAW file will be unreadable or cannot be used to reproduce the photographer’s original interpretation.”⁸

Respondents to the survey identified *Joint Photographic Experts Group (JPEG)* as the most common file format for in-camera capture. Designated an ISO standard in 1994, JPEG has gained industry-wide support regardless of capture device (e.g., camera, cell phone, pads etc.). As an open standard, the JPEG specification is made available to the public, and it is cross-platform operable, which means that the image and its metadata should remain intact when they are transmitted across systems and software applications. The drawback to the format is its use of lossy compression, which enables the file to be transmitted quickly (making it ideal for email attachments and social media sites), but results in loss of information each time the file is saved. Generational loss due to re-compression results in degradation of the image quality, making the JPEG format insufficient for preservation purposes. Thus, the survey findings in response to research question (1) reveal that photographers produce images that fulfill their creative and business needs, while providing the opportunity for future re-use. The original born-digital image is equated with the in-camera capture format, which in the case of this survey, is either RAW or JPEG, formats that have known risks for preservation.

During the procedure of creation, technical and descriptive metadata are attached to and/or embedded within the digital image. Technical metadata refer to the settings that are automatically recorded by the capture device (i.e., camera), such as pixel width and height of the image, colour space, and image compression. Technical metadata are used in determining how the image is constructed and the parameters for its digital representation. Survey respondents’ comments regarding the variety of information recorded about digital images, identified their knowledge and use of the *Exchangeable Image File Format (Exif)* for digital still image metadata, which is a specification that was launched by the *Japan Electronics Industry Technological Association (JEITA)* in 1998 and is now backed jointly with the *Camera & Imaging Products Association (CIPA)* to encourage interoperability between imaging devices.⁹ As long as hardware and software support the information model promoted by Exif, the technical metadata are properly exchanged and retained along with the digital image.

⁷ Adobe Systems Incorporated, “In depth: Digital Negative (DNG),” 2012, <http://www.adobe.com/ca/products/photoshop/extend.displayTab2.html>. The Digital Negative (DNG) Specification v1.3, 2009.

http://www.images.adobe.com/www.adobe.com/content/dam/Adobe/en/products/photoshop/pdfs/dng_spec.pdf.

⁸ OpenRAW, “The Problem with Proprietary RAW files,” 2006, <http://www.openraw.org/info/index.html>.

⁹ Standardization Committee, “Exchangeable image file format for digital still cameras: Exif Version 2.3,” CIPA DC-008-2010/ JEITA CP-3451B, April 26, 2010, http://www.cipa.jp/english/hyoujunka/kikaku/pdf/DC-008-2010_E.pdf.

The drawbacks to the use of Exif metadata standard are that not all of its elements are mandatory, which means that not all devices or software are required to write and read the majority of Exif tags, and not all social media sites preserve image metadata after upload.¹⁰ The lack of universal write/read support for all Exif tags means that critical system metadata that can assist in uniquely identifying the born digital image or provide assurance of data integrity in the future may be stripped from the image during routine procedures over creation and use. Analysis of the Exif schema in terms of its record-keeping capabilities shows that it fails to provide information about digital image context, hierarchical information about relationships between images in an aggregation of images, and processing history about the image. The schema describes the technical aspects of the image itself and the capture device, but it does not give contextual information regarding external agents such as photographer name, or the business activities or management processes it supports.

The most common type of information that survey respondents record about their digital images is descriptive. The descriptive information identifies the context of image creation (i.e., who created the image, when and where it was taken, and why), and explains the content of the image (i.e., persons, locations, and subject matter represented in the image), for purposes of access and retrieval. Essentially, descriptive metadata are explanatory notes that photographers add to active images, with the aid of commercially available software, in order to identify the persons, actions, and matters related to image creation and use. The value of descriptive metadata is its capacity for establishing record identity; however, metadata must remain persistently linked and be managed along with the image to ensure maintenance of authenticity. Since 1991, the *International Press Telecommunications Council (IPTC)* has maintained a metadata standard to transfer a data object, which may be an image file or a combination of text and image, along with its pertinent information, such as creator's name, location, subject matter, and copyright/usage notice, between systems. In the past ten years IPTC has worked with Adobe Systems Inc., and professional photographers to devise an XML based metadata schema (IPTC Core) that offers photographers a reliable and convenient method of applying descriptive metadata to their digital images using templates made available in Adobe software products. In July 2010 IPTC released a comprehensive document entitled "Photo Metadata: Core 1.1 and Extension 1.1", which presents the IPTC Core specification and the more recent IPTC Extension specification—a combination of descriptive and rights based metadata.¹¹ Over ninety-percent of survey respondents believe it is important that their images can be proven to be theirs and are properly credited to them. The addition of metadata is one of the methods photographers use to protect their digital images. By implementing a standardized metadata profile via image management software a control is exercised over the procedure of creation and use, which greatly assists in establishing the identity and integrity of born-digital image files.

Unlike the Exif schema, the IPTC metadata schema is not read-only but has dynamic fields of information that may be changed throughout the lifecycle of the image, depending upon the management

¹⁰ David Riecks, "Social media Photo Metadata use Survey," 2009-2012. Controlled Vocabulary website, <https://spreadsheets.google.com/pub?key=tceeiYNw8ZDC0N52UgRcgnA&single=true&gid=0&output=html>.

¹¹ IPTC, "IPTC Standard Photo Metadata: IPTC Core 1.1/ IPTC Extension 1.1, July 2010, http://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata-201007_1.pdf; The rights based metadata is structured according to the *Picture Licensing Universal System, (PLUS)* developed in 2006 for worldwide use. PLUS recognizes the importance of standardized image metadata to ensure the long-term preservation of image content and context across networked systems. The initiative works closely with *Digital Object Identifier (DOI)*, IPTC and Adobe Systems Inc., see The PLUS Coalition, "Picture Licensing Universal System," 2011, <http://www.useplus.com/index.asp>.

and use of the image. The schema provides photographers with a method of capturing attributes of the digital image's creation and use that contribute to uniquely identifying the image; therefore, it is more effective than Exif for the classification and retrieval of digital images in collections. The IPTC metadata schema does not have the capacity to present hierarchical levels and relationships within an aggregate, or an image processing history that would provide a greater understanding of the changes made to an image throughout its lifecycle. Thus, in response to research question (2) the survey findings reveal that photographers are actively engaged with capturing technical and descriptive metadata in an effort to ensure the identity and integrity of their images. The majority of respondents intend to have their images accurately displayed and credited.

7. Preservation and Transmission

The majority of survey respondents is concerned with the longevity of their digital images and incorporates a procedure for long-term storage into their workflow. Image preservation activities include (in order of frequency): selecting storage media (i.e., CD-R, DVD-R, and external drives), designating file format for originals and surrogates (i.e., RAW, TIFF and JPEG), unique file naming that identifies relationships between originals and surrogates, and using software and capture hardware with specific attributes (i.e., batch metadata capabilities and cataloguing offline CDs.) Survey respondents described procedures for transferring and/or copying in-camera images (i.e., originals) to CDs and DVDs immediately following a shoot and then creating digital surrogates to function as working files which undergo edit operations. Comments made by respondents express concern with the longevity of optical storage and seek advice regarding the best brand of "archival" CD and DVD.

The survey findings show that the measures photographers currently take to protect their image files involve making back-ups and refreshing optical storage media by making 'read-only' copies of CDs and DVDs on a regular basis. No mention was made by respondents of providing file verification after transfer processes.¹² The practice of migrating older image file formats is less common. Protecting digital images from loss and corruption due to technological obsolescence and media fragility is only one part of a preservation strategy. Activities aimed at protecting digital images from unauthorized access and destruction assist in ensuring integrity. The survey findings regarding security measures to protect digital images held within systems and stored on removable storage media show that less than half of respondents apply any type of security measures. The transmission of born digital images outside the personal workspace to facilitate client review, assignment submission, and personal promotion present an opportunity for unauthorized access. More than half of survey respondents present their images on the World Wide Web, and of that group, the majority manages access using a custom built database or a vendor management package. In 2009 a study into the preservation of photo metadata by social media websites was initiated to determine the degree of image metadata support provided by social media sites

¹² In early 2010, tools for file verification using MD5 checksums started being discussed by members of the photographic community interested in digital image preservation. Reasons for file verification include ensuring data integrity of transfers and integrity of stored collections on removable media. David Riecks, "The Trouble Transporting Tribbles (or File Verification using MD5 Checksums)," July 2010, <http://www.controlledvocabulary.com/imagedatabases/file-verification.html>.

and services including Facebook, Flickr, Twitter and Blogger.¹³ Preliminary results of the study reveal inconsistencies in the way image metadata are handled; however, routine tasks such as uploading and resizing digital images were found to remove both Exif and IPTC metadata from the digital image. Thus, in response to research questions (2) and (3) the survey findings reveal that photographers use a variety of commercial software and storage media to manage and preserve their digital image collections. Survey respondents are willing to use products that adhere to digital preservation standards; yet, many photographers shape their digital practice to meet the growing demands of clients for faster turnaround times and more versatile images, which inevitably results in adopting new technologies.

8. Future Directions

The Survey provided a unique opportunity to engage with a specific “creator community” at a key point in their transition from analogue to digital practice. As a result, we gained valuable insight into procedures conducted by creators throughout the creation, use and preservation of born-digital images, which led us to re-evaluate the role and responsibilities of the creator in the digital environment. Ongoing efforts by the photographic community and imaging industry to increase awareness and support for open formats, standardized image metadata and interoperability across networks contribute to the long-term preservation of born-digital images as reliable and authentic records.

The survey findings are relevant to archivists and cultural heritage professionals that are responsible for preserving and creating access to born-digital images. As individual photographers, cultural organizations and government agencies continue to embrace digital practices, the result will be an increase in born-digital accessions to archival repositories. In the past three years, the *United States National Archives* has accessioned over 700,000 digital images (in addition to analogue photography) and of that amount, approximately 100,000 of them are already available through their Online Public Access (OPA) system.¹⁴ There is evidence of the growing number of born-digital archives everywhere, for example the City of Vancouver hosted the 2010 Olympic and Paralympic Winter Games and as a result, the Vancouver City Archives acquired twenty-five terabytes (25 TB) of born-digital materials including videos, images and office files.¹⁵ In the traditional paper-based environment archivists often had years to accession (i.e., the physical and legal act of transferring materials from a donor to an archival repository) and process (i.e., the physical and intellectual activities of arranging and describing) the archival materials; however, the fragile nature of digital media and the expectations of scholars to have immediate access to digital archives combine to place immense pressure on repositories. As archives acquire born-digital and hybrid collections many of the challenges to preserve these materials and make them available to the public provide an opportunity to re-visit archival theory and practice.

Since the work begun by InterPARES, archivists and cultural heritage professionals have conducted research and executed pilot projects aimed at understanding the complexity of born-digital collections. Recent publications including *Digital Forensics and Born-Digital Content in Cultural Heritage*

¹³ David Riecks, “The Controlled Vocabulary Survey regarding the Preservation of Photo Metadata by Social Media Websites,” 2009-2012. Controlled Vocabulary website, <http://www.controlledvocabulary.com/socialmedia/index.html>.

¹⁴ Wade, “From Analog to Digital.”

¹⁵ Courtney C. Mumma, Glenn Dingwall, and Sue Bigelow, “A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Gamers Fonds: or, SELECT * FROM VANOC_Records AS Archives WHERE Value='true',” *Archivaria* 72 (Fall 2011): 93-122.

Collections and findings of the *Born Digital Collections: An Inter-Institutional Model for Stewardship (AIMS)* project (2009-2011) present the preservation community with valuable case studies conducted in a variety of organizational and institutional settings, which significantly extend the research products created by earlier studies such as the *Personal Archives Accessible in Digital Media (Paradigm)* project (2005-2007).¹⁶ Oddly absent from these studies is a direct discussion about born-digital image collections and the role of photo metadata to support archival activities including acquisition, arrangement and description, preservation and access.¹⁷ As demonstrated in this paper, photo metadata contribute to the identity and integrity of born-digital images; therefore, a better understanding of self-describing digital images within the context of archival practice could contribute to current projects being conducted within the archival community on visual literacy, online discovery tools, digital records forensics, and digital preservation systems.

¹⁶ Mathew Kirschenbaum, Richard Ovenden, and Gabriela Redwine, "Digital Forensics and Born-Digital Content in Cultural Heritage Collections," *Council on Library and Information Resources*, 2010, <http://www.clir.org/pubs/abstract/reports/pub149>; AIMS, "AIMS-Born Digital Collections: An Inter-Institutional Model for Stewardship," January 2012, http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf; Paradigm, "Project Overview," 2008, <http://www.paradigm.ac.uk/about/index.html>. Also see Susan Thomas, "A Practical Approach to Preservation of Personal Digital Archives," Final Report to the Joint Information Systems Committee, March 2007, v1.0, *Paradigm Project*, <http://www.paradigm.ac.uk/projectdocs/jiscreports/index.html>.

¹⁷ The Digital Images Archiving Study, published in 2006 conducted by the Arts and Humanities Data Service (AHDS) and funded by the Joint Information Systems Committee (JISC) of the Higher and Further Education Funding Councils and the Arts and Humanities Research Council explored digital image content, user expectations and a life-cycle model for preservation. In the final report, the role of metadata (technical, administrative and discovery) is highlighted as an area for further research and development, especially in light of the different metadata standards supported by the digital photography industry (i.e., photojournalists, professional imaging software and hardware, online photo sharing networks) and the standards supported by cultural heritage institutions for the purposes of management and preservation.

Essential Skills for Digital Preservation

Addressing the Training Needs of Staff in Small Heritage Institutions

Angelina Altobellis

Northeast Document Conservation Center

Abstract

Since 1995, the Northeast Document Conservation Center's "Digital Directions" conference (formerly "School for Scanning") has trained nearly 5,000 cultural heritage professionals worldwide in best practices for creating and managing digital collections. The breadth of topics covered over the past 17 years demonstrates the growth of our ability to create, share, and preserve digital assets. Yet as bits overtake atoms in the construction of our cultural record, it is becoming clear that digital preservation training cannot take a "one size fits all" approach. There is a need for training that takes into account the staffing and funding limitations of small institutions.

Author

Angelina Altobellis is Preservation Specialist for the Northeast Document Conservation Center (NEDCC) in Andover, Massachusetts. She performs preservation needs assessments of analogue and digital heritage collections, consults on long-range preservation planning, and teaches webinars on the preservation of analogue and digital collections. Angelina is a member of the Society of American Archivists and the American Library Association. She holds an M.L.I.S. from Simmons College, an M.A. in Comparative Literature from the University of Texas-Austin, and a B.A. in Art History from the University of Massachusetts-Amherst.

1. Introduction

Since 1995, the Northeast Document Conservation Center's "Digital Directions" conference (formerly "School for Scanning") has trained nearly 5,000 cultural heritage professionals worldwide on best practices for creating and managing digital collections. What began as a one-day conference on the technical aspects of scanning has grown to a three-day event designed to educate participants on managing the full life-cycle of digital objects, from creation, to curation, to use. It has been presented in a dozen cities in the United States, and internationally in Cuba (2001) and the Netherlands (2002).

The breadth of topics covered over the past 17 years demonstrates the popularity of digitization, as well as the proliferation of born-digital collections, and the tremendous growth of our ability to create, share, and preserve digital assets. At the same time, feedback from conference participants has revealed a disparity in training needs. Where staff from large, well-funded institutions are eager for information on topics such as advanced metadata and data curation, staff from small or under-funded institutions seek practical strategies to create and manage digital collections "on a shoestring."

Based on NEDCC's conference evaluations and survey data, this paper will discuss the digital preservation training needs of small or under-funded collecting institutions. It will then discuss recent developments by the Center to better address the training needs of these institutions through a revamping of its Digital Directions conference into two tracks and the development of an online training series. The goal of this paper is to initiate fruitful discussion on strategies for closing the "digital preservation training gap."

2. From School for Scanning to Digital Directions

NEDCC was established in 1973 to provide cultural heritage institutions with expertise in the preservation and conservation of paper-based materials. From the beginning, its services have been geared toward institutions lacking the resources—whether equipment, funding, or staff—to maintain preservation or conservation programs in-house. Education and training are vital to NEDCC’s mission. Since the establishment of its Preservation Services unit in 1980 (formerly Field Services), NEDCC has presented more than 1,400 workshops, conferences, and lectures. These offerings routinely gather audiences from across professional lines to foster discussion and encourage networking.

Digitization of cultural heritage materials by libraries and museums had been underway for roughly a decade when NEDCC held its inaugural School for Scanning on April 13, 1995 at the John F. Kennedy Library in Boston.¹ Ann Russell, then NEDCC’s executive director, has recalled that “most of the audience members at the 1995 conference had no first-hand experience with digitizing collections materials. Their concerns were about how to get started, and their most pressing question was whether or not they should digitize.”² By and large, digitization was the province of major research libraries and federal agencies (two of the presenters at the first School for Scanning came from Cornell and Yale); for everyone else, its future was still somewhat uncertain. To this point, the conference brochure questioned, “Will digitization...become a tool of the preservation community? Is digital preservation, in fact, already a reality?”³

Attendance at the first School for Scanning topped 300 people, persuasively suggesting that there was, in fact, keen interest in making digitization a tool of the preservation community.

With startup funding from the Andrew W. Mellon Foundation and subsequent support from the National Endowment for the Humanities, NEDCC delivered the conference twelve times over the next twelve years to sold-out crowds. It sought to educate a broad audience with a “deliberate focus on decision making, as opposed to recommending specific products and procedures.”⁴ The goal of this approach was to provide each participant with a grounding in core concepts to inform local implementation of digital projects and programs. Topics presented regularly included content selection, rights management, digitization of text and images, quality control, costs, and metadata.

By the turn of the 21st century, numerous institutions had launched, or were planning to launch, ambitious digitization projects. While these focused most often on collections at larger institutions, collaborative efforts such as the Colorado Digitization Project began to emerge as well, sometimes extending digitization’s reach to smaller ones.⁵ In the years that followed, a widening range of cultural heritage professionals began building or contributing to digital collections, gaining familiarity with the

¹ Leslie Johnston, “Who Do You Want to Be Today?” *The Signal* (blog), August 28, 2012, <http://blogs.loc.gov/digitalpreservation/2012/08/who-do-you-want-to-be-today/>. Johnston describes working on “digitization and metadata creation and normalization” in the mid-1980s.

² Ann Russell, “Training Professionals to Preserve Digital Heritage: The School for Scanning,” *Library Trends* 56, no. 1 (2007): 289.

³ Northeast Document Conservation Center, *School for Scanning: A Conference on Digitization, Microfilm, and Preservation*. (Andover: NEDCC, 1995.) Brochure.

⁴ Russell, “Training Professionals to Preserve Digital Heritage,” p. 291.

⁵ Nancy Allen, “Collaboration through the Colorado Digitization Project,” *First Monday* 5, no. 6 (June 5, 2000), <http://firstmonday.org/article/view/755/664>.

associated technology and requirements for digital projects. At the same time, many others remained at the beginning of the learning curve, hoping to start their own digital projects but unsure how.

Both groups turned to School for Scanning to deepen their understanding of the digital landscape. As Ann Russell observed, expectations of the conference became divided.⁶ One group sought to build on or reinforce existing knowledge, and to learn about advanced topics and emerging trends that might inform their own digital programs. The other group sought an introductory survey that would enable them to join the conversation. “I think we might be farther along than most in the audience,” noted one participant in her evaluation of the 2005 conference in Boston. “But I realize that as I am disappointed about not learning enough, the audience seemed lost halfway through the second day. Maybe two tracks? A learning track and an advanced track?” Other comments stressed the need to make presentations more helpful to staff at smaller institutions:

- “It might be helpful (for those from smaller institutions) to have speakers from small to mid-sized museums/archives who have successfully implemented a project.” (Los Angeles 2003)
- “The message [that] there is no one best solution was both comforting and frustrating. I find it requires some translation to make it practical/useful for small institutions.” (Chicago 2004)
- “I may have learned that our small institution does not have the resources to even attempt digital projects. It’s too big, too overwhelming, and way too complex. It is discouraging to know what we need to do, but cannot possibly do given our resources.” (Boston 2005)

Clearly, NEDCC needed to find a way to bridge a growing gap. Demand for training opportunities like those offered at School for Scanning remained strong, as evidenced by repeatedly high attendance levels: 2005 marked a record with 429 participants. Yet feedback plainly revealed that a different approach was needed for the conference to remain useful to professionals at different skill levels and from different institutional settings. Moreover, the thinking about the nature of digital collections was changing. Collection managers were grappling with strategies for managing not only the products of past digitization projects, but also an influx born-digital materials. Discourse developed around questions of digital continuity and the lifecycle of “digital objects.” It was time for School for Scanning to evolve in both approach and scope.

In June 2008, NEDCC launched Digital Directions—billed as “the new School for Scanning”—in Jacksonville, Florida. The name reflected both a new structure and an expanded scope. Concurrent breakout sessions were introduced to allow participants to choose their own “direction.” This also resulted in smaller sessions, creating a more comfortable, classroom-like environment for lively discussion. Content grew to encompass the full lifecycle of digital objects, from planning to creation to sustainability, and dual focus was placed on teaching both conceptual ideas and practical strategies. Core topics such as metadata, content selection, rights management, project planning, and digitization of various formats remained, but they were complemented by sessions on the basics of scanning, equipment selection, preparing materials for digitization, standards and best practices for sustainability, and digital disaster planning.

⁶ Russell, “Training Professionals to Preserve Digital Heritage,” p. 295.

3. Digital Directions 2012

NEDCC kicked off planning for its third Digital Directions conference in 2011. The planning phase began with a survey to understand the profile of the potential audience and identify specific training needs. Paper copies of the survey, with a URL to the online version, were distributed at the Center's exhibit hall booth at the Society of American Archivists annual conference in Chicago in August 2011. Subsequently, announcements with a link to the survey online were emailed to SAA attendees on September 1, and to NEDCC's entire email list on September 28.⁷ A link to the survey remained on NEDCC's home page through November 2011.

Respondents were asked to identify their institution type, their roles, responsibilities pertaining to digital collections, the makeup of their digital collections, and training interests. A total of 330 people completed the survey. Not surprisingly given its distribution, the largest number of respondents—63%—identified themselves as archivists. They represented educational institutions (39%), archives (36%), libraries (34%), non-profit organizations (21%), government agencies (15%), museums (13%), historical societies (10%), corporations (4%), and the self-employed (4%). (The categories for this question were not mutually exclusive.) The vast majority of respondents reported that the bulk of their current and planned digital collections were products of digitization (82%), while just 18% reported that the bulk comprised born-digital collections. This figure is interesting in light of the fact that 64% were interested in learning more about born-digital objects and collections, strongly suggesting that they anticipate—or would like to lay the groundwork for—growth in this area. Half or more of respondents were also interested in learning more about standards and best practices (73%), user interfaces and access tools (58%), storage (56%), workflows and management software (55%), metadata basics and tools (54%), project planning (52%), reformatting strategies and equipment (51%), and strategies for partnerships and collaboration (50%).

A comment box was provided for respondents to share “additional areas of need or ideas for Digital Directions 2012, as well as for NEDCC's educational offerings, both in-person and Web-based.” The 73 responses received were enlightening. Several respondents explicitly identified themselves as coming from small institutions while emphasizing a need for information on practical tools and strategies in areas including software, storage, project planning, digitization, and digital asset management:

- “Best Practice for storage and accessibility for small institutions. Developing a volunteer support program for digitizing collections.”
- “Many digital project workshops, planning documents and standards are designed for organizations with funding sources, trained staff dedicated to the project, and very ‘best of all possible worlds’ to those of us in minimally staffed and small organizations. Many have guidelines that are just not realistic for a small institution with primarily student workers and minimal funding. Would like more ‘real world’ and viable solutions for very small, underfunded organizations—along the lines of ‘this is what you can do with little money and staff.’”
- “I have major issues with open-source materials because of the lack of [IT] support. I’d love something like the SAA Session—Practical Approaches to Born Digital—What works and what doesn’t—for the small shop. A webinar would be perfect.”

⁷ The September 1 and September 28 emails were sent to 1,560 and 3,012 people, respectively.

- “Smaller LAM [library-archive-museum] institutions are (or would be) very interested in sessions or offerings focusing on **practical** and user-friendly tools for digitization, digital preservation, and digital management. Many offerings tend to focus on tools, strategies, or frameworks beyond the budgetary or IT means of smaller institutions.”

Other suggested topics included advanced metadata, workflow management, digital repositories, data curation, open-source software, capture practices for different formats, and communicating with IT staff.

NEDCC’s third Digital Directions conference took place in Boston from June 13-15, 2012. The audience of 172 people represented 31 U.S. states and the District of Columbia, and four Canadian provinces. Roughly half represented academic libraries; 10% represented historical societies; 15% government agencies; and 13% museums. The remainder represented non-profit institutions, corporations, students, and consultants. Close to one quarter (24%) of the audience at Digital Directions 2012 represented institutions with more than 50 FTEs; slightly more (28%) represented institutions with 5 FTEs and fewer.

Foundational sessions on the first day covered the conference’s central theme of “Creation, Curation, Use,” and laid the groundwork for concurrent breakout sessions on the second day. As with previous Digital Directions conferences, participants could develop their own track based on their training needs. One track focused primarily on practical aspects of digital collection building, including tools and techniques for digitization and building a digital production lab. The other focused primarily on strategies for planning and administration of digital projects and programs. The final day brought all participants back together for sessions focused on collaboration and partnership building.

Attendance averages for both tracks were essentially equal, a fact that is especially interesting given that more conference participants identified as practitioners than administrators. This suggests that professional staff who work with digital collections do so wearing multiple hats—and must therefore possess or develop a range of skills. In their conference evaluations, participants were asked to indicate which of the topics covered in the conference would be most useful to them. By far the most popular choices were digital creation (63%), digital curation (60%), and metadata (60%).

Participant feedback was positive overall, with 91% reporting that they learned as much or more than they expected. There was also promising data on the conference’s impact on the development and management of digital collections: 80% of participants stated that they were likely or very likely to improve an existing project or program as a result of the conference; 73% would set new priorities for work; and 57% were motivated to plan a new project. Comments suggested that Digital Directions 2012 struck a good balance between theory and practice. Several praised the conference’s relevance to smaller institutions. An archivist at a small academic library offered a particularly encouraging perspective:

Although I’m a “lone arranger” working on a very part-time schedule, I am inspired to at least create a plan towards a digital program and gathering support by building partnerships with other departments and institutions! Through this conference, I feel equipped to begin pursuing projects as best I can, with the resources I have.

Comments also raised two main concerns, however. Several participants from museums felt that the conference was too focused on the needs of libraries. Others felt that there was too much content packed into the three days, sometimes finding it difficult to choose between concurrent sessions. These are issues for NEDCC to explore in future program development.

4. Fundamentals of Digitization

Another concern is that multi-day conferences are often cost-prohibitive and too time-intensive for the smallest institutions, some of which are volunteer-run or very minimally staffed. In a statewide preservation needs survey conducted in 2010 by NEDCC and the Massachusetts Board of Library Commissioners, just over one-quarter of all respondents had no full-time staff.⁸ Nearly half of these were historical societies; just 2% were academic libraries. A comparison of data on training preferences points emphatically to a need for varied approaches to training. When Digital Directions 2012 participants were asked their preferred format for future training, 73% selected “multi-day, in-person conferences/workshops.” By sharp contrast, those with no full-time staff responding to the 2010 Massachusetts survey overwhelmingly preferred either half-day or full-day training programs (70%).

NEDCC is working to meet the needs of these institutions by expanding its online training offerings. In September 2011, it launched a “Fundamentals of Digitization” webinar series, composed of two-hour webinars on the basics of building sustainable digital collections. Participants can register for as few or as many webinars as needed; topics covered include surveying digital preservation needs, selection for digitization, metadata, digital project planning, copyright, workflow for digital projects, and digital disaster planning. The webinar format allows institutions to participate in training from their homes or offices, eliminating the cost of travel. The series is now beginning its second year.

5. Conclusion

As bits overtake atoms in the construction of our cultural record, it is becoming clear that training to preserve those bits cannot take the form of a “one size fits all” approach. There is a need for digital preservation training grounded in standards and best practices that also takes into account small heritage institutions’ “trilemma” of lacking human resources, lacking funds, and lacking technical skills.”⁹ What is at stake? In his 2010 article “Preservation in the Age of Google,” Paul Conway warned that “In the age of Google, nondigital content does not exist, and digital content with no impact is unlikely to survive.”¹⁰ Scholarship is ever more oriented toward the digital realm. Collecting institutions of all sizes must be able to contribute their holdings, whether physical or born-digital, to the networked environment, and they must be able to knowledgeably manage their digital holdings to ensure their longevity. To be sure, this will require more than training alone. Shared infrastructures and collaboration will be essential. But to be effective participants in the cultural heritage and research communities of the future, those charged with stewarding the human record must understand how to build and maintain good digital collections.¹¹

⁸ Northeast Document Conservation Center and the Massachusetts Board of Library Commissioners, *Connecting to Collections: A Statewide Preservation Survey, Final Report* (Andover, MA: Northeast Document Conservation Center, 2011): 10, <http://mblic.state.ma.us/advisory/preservation/c2c.pdf>.

⁹ Guntram Geser and Andrea Mulrenin, “Are Small Heritage Institutions Ready for E-culture?” (paper presented at the 2004 International Cultural Heritage Informatics Meeting, Berlin, August-September 2004), 5, http://www.archimuse.com/publishing/ichim04/3687_GeserMulrenin.pdf.

¹⁰ Paul Conway, “Preservation in the Age of Google,” *Library Quarterly* 80, no. 1 (2010): 64, <http://hdl.handle.net/2027.42/85223>.

¹¹ “A Framework of Guidance for Building Good Digital Collections,” National Information Standards Organization, accessed September 2, 2012, <http://framework.niso.org/node/5>.

Small and Large Scale Digitization: Towards a Shared Conceptual Model

Building Sustainable Digital Cultural Heritage Collections

Towards Best Practices for Small-scale Digital Projects

Peter Botticelli, Patricia Montiel-Overall and Ann Clark

University of Arizona, School of Information Resources and Library Science

Abstract

In the past decade, many institutions have undertaken small-scale digital projects (often involving a few hundred items) designed to expand access to cultural heritage online. This work examines a number of theoretical and practical challenges involved in planning and carrying out small-scale digitization projects involving culturally sensitive artefacts. This work builds on a study, Building Digital Cultural Heritage Collections in Arizona, funded by the U.S. Institute of Museum and Library Services. Our larger aim is to help define best practices for online documentation of historically under-represented communities.

Authors

Peter Botticelli is Assistant Professor of Practice in the University of Arizona School of Information Resources and Library Science, where he directs the Digital Information Management (DigIn) and Archival Studies certificate programs. He has a Ph.D. from the University of Illinois in history, and a Master's degree in Archives and Records Management from the University of Michigan. He teaches courses on digital curation, scholarly communication, and digital preservation, and has published research in history and archival studies.

Patricia Montiel-Overall is Associate Professor in the University of Arizona School of Information Resources and Library Science. She has a Ph.D. from Stanford University in education, and an M.A. in Library and Information Science. Her research interests include equity of access for diverse populations and cultural competence particularly as they relate to school and public librarianship, and to library services for Latino and Native American populations.

Ann Clark is a 2012 M.A. graduate from the University of Arizona School of Information Resources and Library Science. As a graduate assistant, she was responsible for a research project entitled "Building Digital Cultural Heritage Collections in Arizona" funded by the Institute for Museum and Library Services. Ann also completed the United Nations Secretariat's 2011 Summer Internship Programme, in which time she worked on digital records management strategies including migration planning, file classification scheme editing, and retention schedule analysis for the Office of Information and Communication Technology's Knowledge Management Services Section.

1. Introduction

In recent years, cultural heritage professionals have paid much attention to the sustainability of digitization activities and Web-based exhibits and repositories.¹ Within the digital preservation

¹ In the U.S., the Council on Library and Information Resources has published a number of influential reports touching on the theme of sustainability in digitization. See, for instance, Diane M. Zorich, *A Survey of Digital Cultural Heritage Initiatives and Their Sustainability Concerns* (Washington, DC: CLIR, 2003). See also: *Access in the Future Tense* (Washington, DC: CLIR, 2004), and *No Brief Candle: Reconceiving Research Libraries for the 21st Century* (Washington, DC: CLIR, 2008). See also, Alison Babeu, "Rome Wasn't Digitized in a Day": *Building a Cyberinfrastructure for Digital Classics* (Washington, DC: CLIR, 2011), 242-248.

community, the problem of sustainability has increasingly been linked to access concerns as well as to solving the underlying problem of technology obsolescence.² More generally, digitization of cultural heritage materials has increasingly been viewed by libraries, archives, and museums as a core function—as opposed to a peripheral or experimental activity—enabling institutions to engage users more effectively online, and ultimately to add value to non-digital as well as born-digital collections. In this regard, sustainability of digital programs depends on institutions building a robust Web presence that makes collection items more discoverable as well as accessible online.

In this context, a key objective of libraries since the 1990s has been to increase the scale of digitization, especially in the scanning of books and journals, and to build large aggregations of digitized collection items. Thus far, the results of mass digitization efforts have been mixed, at least partly due to copyright issues.³ A number of large-scale projects have been successful in aggregating public domain works accessible on the Web,⁴ and yet, the digitization of archives and special collections especially has continued to be characterized by relatively small projects, typically involving hundreds or thousands of items (as opposed to hundreds of thousands or millions), often organized on a non-routine or episodic basis by repositories.

In this context, our work focuses on contributions and potential contributions to the Arizona Memory Project (AMP), an online repository maintained by the Arizona State Library, Archives and Public Records.⁵ AMP's mission is to support local digitization efforts by a wide range of libraries, museums and historical societies in the state, especially by hosting the resulting collections through a central CONTENTdm repository. Since AMP was founded in 2006, 64 different institutions have contributed over 140 distinct collections, representing about 90,000 total digital objects. With collections averaging less than 700 objects each, AMP serves as a useful case example of how small-scale digital projects are currently being aggregated and made discoverable online in the U.S. today, with many states hosting repositories comparable to AMP.⁶

As we will see, digital repositories such as AMP offer abundant evidence as to the potential impact of digitization, and the sharp limitations faced by cultural heritage institutions in developing digital assets. In general, the small-scale digital projects we see in AMP represent a broad first step—through projects often organized on an experimental basis and with grant funding—toward building the sustainable institutional capacity needed to fully document cultural heritage in digital form. Having taken initial steps in digitization, as represented by AMP and similar efforts, it's becoming increasingly important to closely evaluate the

² Blue Ribbon Task Force on Sustainable Digital Preservation and Access, *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information* (February 2010), accessed 31 Aug. 2012, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

³ With the Google Books project having digitized over ten million items, and with access stymied by intellectual property rights concerns, we might conclude that the cause of mass digitization has reached a plateau, although Google itself estimates that the total number of books in existence might be as high as 130 million, accessed 31 Aug. 2012, <http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html>.

⁴ In the U.S., for instance, projects such as JSTOR, American Memory, and HathiTrust can all be viewed as transformative for library services.

⁵ <http://azmemory.azlibrary.gov/cdm/about> (accessed 31 Aug. 2012).

⁶ It might be pointed out that in spite of the Library of Congress's status as the de facto national library of the U.S., it does not have a legislative mandate to aggregate and provide access to collections not owned by the Library itself, thereby limiting the potential for a digital repository at the federal level. Hence, a large portion of digitization across the U.S. has been funded and organized at the state level.

outcomes of the first generation of small-scale digital projects, both regarding the cost of digitizing items, and especially measuring the value added to collections through Web exhibits and repositories.

In Arizona and the Southwestern U.S., digitization efforts face major challenges in documenting a diverse cultural landscape with an often contentious history shaped over centuries by long-distance migration, scarce natural resources (especially water), frequent territorial conflicts, and the oppression of indigenous cultures. Today, the much-travelled border between Arizona and Mexico is charged politically, making it all the more important for information professionals seeking to digitize and expose cultural heritage collections online to have clear guidance on best practices for digital collection building and access.

In evaluating the small-scale digital projects in AMP, our first concern is with appraisal—the choice of artefacts to digitize and include in AMP. Here it's essential that we consider resource constraints that have sharply limited the scale of digital projects in Arizona, as well as the range of artefacts that could potentially be digitized, thereby filling important gaps in the state's social memory—particularly involving non-English speaking cultures. In addition, many institutions in the region face complex issues regarding the ownership and management of collections, particularly involving indigenous artefacts. As we will see below, the experience of repositories in Arizona to date suggests that substantial research and collaboration with local communities is needed to appraise artefacts for digitization while upholding the spirit of the *Protocols for Native American Archival Materials* and similar guidelines for respecting indigenous knowledge.⁷

In an effort to understand how digitization efforts in Arizona might be expanded and made more culturally responsive, we began this project with a series of case studies funded by the U.S. Institute of Museum and Library Services (IMLS) and conducted in partnership with Arizona State Library, Archives and Public Records. Our aim in developing the case studies was to investigate economic, professional and cultural factors affecting repositories' decisions on whether or not to digitize their collections. Through semi-structured interviews, we initially investigated a sampling of cultural heritage institutions in the Tucson area, seeking data on their current collection development and preservation plans, organizational infrastructure, organizational culture regarding attitude towards technology, policies on information management and access, technology infrastructure, and scale and technical characteristics of collections that might be digitized in the future.

Our initial findings showed that many institutions in the region have not yet undertaken digitization projects. This is not due to a lack of interest; rather, we found that many organizations have clear aspirations to digitize, and some have engaged in initial planning for digital projects even if they have thus far been unable to initiate such activities. Among those that have undertaken digital projects, including those who have made contributions to AMP, we found a common pattern of organization that tends to be informal, episodic, and peripheral to the institution's larger mission and core activities. In such an environment, the concept of "best practices" for digitization becomes problematic, as institutions clearly have to strike a difficult balance between what is ideal and what is expedient.

Of course, in the archival context, such trade-offs have long been a factor in organizing and carrying out documentation strategies. This concept, as originally defined by Helen Samuels, was intended to move the archives profession toward a more formal, rigorous approach toward documenting topics of importance to institutions—with the appraisal of records going beyond evidential values and

⁷ <http://www2.nau.edu/libnap-p/> (accessed 31 Aug. 2012).

instead seeking to capture a broader range of information of permanent value.⁸ As Richard Cox and others have noted, by widening the scope of appraisal in an effort to more fully document society, tensions are bound to arise in the process of appraising particular records, especially as the volume of records produced by society has risen exponentially in recent times. Moreover, as a growing number of archival theorists have noted, archivists today face greater demands for accountability and transparency in making appraisal decisions. Whereas Cox argued in the early 1990s that “even a faulty archival appraisal decision or decision process is better than records surviving haphazardly or not surviving at all”⁹—a view most would still agree with today—and yet, today archivists face new pressures to represent society’s collective memory in ways that are more inclusive and context-rich than users might have expected in the past.¹⁰

With the growing critical attention paid by archival theorists to the complexity of social memory as it’s represented in archives, archivists might be tempted toward a skeptical view of documentation strategies—especially in the Web environment. In a 2008 article, Doris Malkmus compares the major challenges faced by formal documentation strategies, the most important of which are resource constraints and the accompanying need to limit the topic to be documented. In general, she finds that broader or more general topics have been less likely to succeed or to be sustainable over time. Likewise, Malkmus argues that successful documentation strategies have usually been “implemented as a series of narrowly focused, sequential projects, rather than as single, comprehensive projects.”¹¹ As Malkmus indicates, archivists have pursued significant numbers of formal documentation strategies in the Web environment, with promising results in some cases. But on the whole, it’s clear that resource constraints have sharply limited the number and scope of formal documentation strategies, and this is certainly the case given the budget constraints facing many archival repositories today.

Judging by the archives literature as a whole, we might conclude that documentation strategies should be regarded as a specialist activity within the archival profession, one that demands a clear focus and a formal institutional mission backed by substantial and ongoing resource commitments, as Cox and others have argued since the 1980s. This longstanding view of documentation strategies might be called into question by the rise of the “More Product Less Process” philosophy,¹² which suggests a way forward for less formal approaches to appraisal as well as archival processing. More work is needed to evaluate and potentially resolve the tensions that might be expected to arise with “minimal processing” as a best practice for appraisal as well as arrangement and description. What is clear is that the rise of the Web has led to a proliferation of what we might term informal or prototype documentation strategies, the results of which make up a large body of digital exhibits available online today, including AMP.

While many digital projects to date might lack the degree of organization and sustained effort we might associate with formal archival documentation strategies, the technical infrastructure built up for first-generation digital projects has much potential to be expanded and refined in future digitization

⁸ Helen Samuels, “Who Controls the Past,” *American Archivist* 49, no. 2 (Spring 1986): 109-124.

⁹ Richard J. Cox, “The Documentation Strategy and Archival Appraisal Principles: A Different Perspective,” *Archivaria* 38 (Fall 1994): 11-36, p. 18.

¹⁰ Terry Cook, “Fashionable Nonsense or Professional Rebirth: Postmodernism and the Practice of Archives,” *Archivaria* 51 (Spring 1994): 14-35.

¹¹ Doris J. Malkmus, “Documentation Strategy: Mastodon or Retro-Success?” *American Archivist* 71, no. 2 (Fall/Winter 2008): 384-409.

¹² Mark A. Greene, “MPLP: It’s Not Just for Processing Anymore,” *American Archivist* 73, no. 1 (Spring/Summer 2010): 175-203.

efforts. Metadata in particular can be seen as an area in which progress has been made, but that also remains a critical bottleneck in advancing digitization. A 2009 survey, for instance, found that over 50 percent of archival collections are not yet discoverable online.¹³ Digitized collections have often been described at a minimal level of detail, forcing users to search for contextual information outside the repository hosting the collection.

The lack of context in Web-based exhibits has been a serious challenge from the beginning, as repositories have a bigger incentive to digitize additional items than to add richer contextual information to their existing Web presence. This is a problem because, as Terry Cook and many others have argued, appraisal has always depended on transparency about the criteria used in selecting records.¹⁴ This is certainly no less true in the digital environment than it was with print records; what is different with digital collections is the additional burden on metadata to reveal the context of individual items. With print records, a user could infer much information about the context of records by their physical arrangement, and in many cases we could rely on collection-level metadata—finding aids and catalog records—to discover items with relevant information.

With Web-based exhibits, however, individual artefacts—usually digitized as a sample or subset of a larger collection—often demand item-level metadata as well as the more traditional types of collection-level description. Hence the problems flagged by Hedstrom, who critiques the way archivists often appear to make “little effort to leave clues about the basis for their appraisal decisions or the contexts in which they are made” in preparing digital exhibits. She also complains about the way much online metadata seems to “pay relatively little attention to the interpretive spin that description places on archival materials.”¹⁵ Our argument is that where context is lacking in Web exhibits, it’s likely to be due to a combination of resource constraints (especially the dependence on short-term grant funding for digital projects) and the experimental or prototypical nature of first-generation digital projects. Hence, as we look forward to second- and third-generation digitization efforts, it’s essential to have a robust conceptual framework we can use to evaluate the results of digital projects, especially in an effort to allocate scarce resources as best we can in adding value to records by providing essential contextual information for users.

In recent years, an influential effort has been made in the library field to apply the idea of “cultural competence” to the management of information resources and services.¹⁶ Cultural competence is a well-established concept in human services such as nursing and social work. In these fields, “cultures” are most commonly understood as patterns of behavior, or as routine social activities that can be observed in a local institution or community setting.¹⁷ By pursuing cultural competence through professional development programs, the goal is to help professionals interact more effectively with people of different backgrounds, partly by acquiring greater knowledge of the cultures they’re serving, but also by giving

¹³ Jackie M. Dooley, “The OCLC Research Survey of Special Collections and Archives,” *Liber Quarterly* 21, no. 1 (November 2011): 125-137, p. 131.

¹⁴ Terry Cook, “Macroappraisal in Theory and Practice: Origins, Characteristics, and Implementation in Canada, 1950-2000,” *Archival Science* 5, no. 2-4 (December 2005): 101-161.

¹⁵ Margaret Hedstrom, “Archives, Memory, and Interfaces with the Past,” *Archival Science* 2, no. 1-2 (March 2002): 21-43, pp. 37-38.

¹⁶ Patricia Montiel Overall, “Cultural Competence: A Conceptual Framework for Library and Information Science Professionals,” *Library Quarterly* 79, no. 2 (2009): 175-204.

¹⁷ Renato Rosaldo, *Culture and Truth: A Remaking of Social Analysis* (Boston: Beacon, 1989).

professionals a deeper understanding of their own cultural background—especially as reflected in the decisions and activities distinct to professions.

Not surprisingly, the idea of culture-as-activity fits closely with the professional values of librarians. Cultural competence also compliments our existing theories of information seeking behavior, while adding a new emphasis on the ways library services can be tailored according to the needs of an increasingly diverse population of users. Montiel- Overall argues the value of cultural competence for information professionals is that it offers a practical framework for addressing information in the social context in which it's created and used. In the American Southwest, the need for culturally competent library resources and services has long been acute, with user needs varying dramatically at the local level.

With respect to archives, the need has long been evident for a culturally competent approach to documentation strategies in the digital environment, particularly as repositories seek to expose records on culturally-sensitive topics involving indigenous knowledge, for instance, or cross-border relations in the American Southwest. Fortunately, such concerns have received much attention by archival theorists in recent years, as the profession has become less focused on its own institutional culture and more attuned to the political and societal dimensions of archival practice. Schwartz and Cook, among others, have argued persuasively that archivists need to deepen their understanding of how society as a whole shapes collective memory, and how archivists act as an integral component of society and not from an impartial or objective position within the political and social environment. Writing in 2002, the authors claimed that by comparison to many academic researchers, archivists have “fallen behind in their theorizing about archives and records, and the power relations embedded in them, shunning the shifting, interactive, and dynamic perspectives of postmodern relativity for the more comfortable and passive stance of the detached observer.”¹⁸ This view sits well with Cook's earlier arguments about the need to treat archival appraisal as a “work of complex scholarship” as opposed to an act of expediency, a “mere process or procedure,” the outcome of which might depend as much on organizational constraints as a conscious analysis of the archival record.¹⁹

Such calls to advance state-of-the-art archival practices are very much in line with the issues addressed by cultural competence. Writing from a library perspective, Montiel- Overall argues that the concept of “information” needs to be broadened beyond traditional (Western) record categories, including “anything that informs, builds, develops, and enriches thinking and human integrative thought.”²⁰ This view is consistent with a growing number of archivists who argue for a broad view of records in collective memory, even as it might appear in forms that lack the permanence we normally associate with paper. Thus, Diana Taylor argues for a view of archives that includes both static artefacts and “repertoire” or “embodied practice/knowledge” as it may appear in social activities such as spoken language, dance, sports, and rituals.²¹

From a cultural competence perspective, there is certainly a need for archivists to base appraisal decisions on a transparent set of criteria, and to shape the institutional mission of the repository around the constituencies we serve, as opposed to the records themselves. Yet the problem in implementing the “postmodern” vision of archives remains a practical one: if we expect archivists to do more analysis in the

¹⁸ Joan M. Schwartz and Terry Cook, “Archives, Records, and Power: The Making of Modern Memory,” *Archival Science* 2(1-2) (March 2002): 1-19, p. 10.

¹⁹ Cook, “Macroappraisal in Theory and Practice,” p. 103.

²⁰ Overall, “Cultural Competence,” p. 182.

²¹ Diana Taylor, *The Archive and the Repertoire: Performing Cultural Memory in the Americas* (Durham: Duke University Press, 2003).

appraisal process for Web-based collections, and especially to create richer contextual metadata, we can't necessarily expect institutions to find the resources needed to support an expansion of digital collections on the Web. In particular, we might expect to see fewer and smaller, yet better described, digital collections available to users, even as we greatly expand the range of potential documentation—both through metadata and access to digitized records—that might be included in online collections.

From our perspective, what stands out is the need for a culturally-competent process for digitization, one that acknowledges the values and constraints of the recordkeeping culture as well as the cultures being documented. It's worth noting that "competence" in this framework is very much a relative and not an absolute value, just as "culture" is a broad term representing many facets of human behavior.

2. Cultural Competence as a framework for evaluating digitization efforts

As noted above, by "culture" we mean a pattern of activity within a distinct social or ethnic group. This definition reaches beyond the idea of cultural heritage as a set of physical artefacts embodying distinct values, beliefs, and traditions, to name a few. It also goes beyond the idea of culture as ethnicity or group identity. By focusing on a broader concept of culture grounded in social activity, our aim is to establish a framework we can use to analyse the work of information professionals as they collect and describe artefacts representing cultures (i.e., the daily activities or events that occur among a group or organization). In other words, to properly evaluate the value of documentation we need to understand the culture that produces artefacts and also the recordkeeping culture that aggregates and preserves information. Our main argument is that the value of documentation strategies of all kinds depends on the reciprocal relationship between creators and collectors. In this regard, it's essential to view cultural competence as a process, not an outcome. In essence, the cultural competence framework developed is designed to help professionals acquire essential knowledge at three basic levels:

2.1 Cognitive dimension

2.1.1 Building cultural self-awareness

In the human services, many researchers have highlighted the need for practitioners to be conscious of their own cultural background as they attempt to communicate with people representing different cultures. In many disciplines, self-examination of customs, values, and social identities is encouraged as a way to help individuals identify actions and beliefs within their own cultures that may prevent effective dialogue or understanding and inhibit cross-cultural interactions. For professionals who work directly with the public, cultural self-awareness has to include both personal or family cultures and the professional cultures in which we operate—and which are often opaque to users. Of course, librarians and archivists have long invested in instructional and reference services in an effort to educate users about the less intuitive aspects of our professional cultures, as reflected in our collections and access systems. But as we'll see in the case of digitized collections and online exhibits, repositories have often provided minimal contextual metadata, both on the meaning of particular items and on the rationale for digitizing certain items and not others. Thus, in designing online exhibits, archivists should not expect users to understand the centrality of provenance in archival arrangement and description. Rather, by understanding the difference in perspective between a provenance-based view of records and the subject-based information needs of many users, we might find better ways to assist users as they search for relevant information.

2.1.2 Acquiring cultural knowledge

Beyond self-reflection, a critical step toward cultural competence involves learning more about other cultures, with the aim of developing a greater sensitivity to cultural nuances. Such knowledge often requires a mix of formal and informal methods of learning. For human services professionals, it's often considered essential to build trusting relationships with member of the community being served, facilitating the sharing of tacit cultural knowledge that might be an essential factor in the outcome of services. Trust is vital in developing a network of confidants who can help mediate between the professional and the community. Here again, reference librarians have a long history of working directly with users, especially face-to-face, to navigate the conceptual and physical obstacles separating users from collections. The problem of assisting users becomes still more complex in the digital environment, as we have to design access systems to support a broad range of information needs, beyond the needs of individual users as might be discerned in a reference interview. Given these complexities, it's helpful to think of cultural competence in digitization less as a function of true expertise (based on formal knowledge) and more as the capacity to engage in a robust dialog with creators and users as we go about designing digital exhibits and exposing metadata online.

2.2 Interpersonal dimension

Within the cultural competence model, the cognitive dimension focuses on the individual, calling for self-reflection and for acquiring knowledge about other cultures. By contrast, the interpersonal dimension involves direct social engagement and dialog, thereby extending and reinforcing the knowledge gained through self-reflection and formal learning. For information professionals, an important task at the interpersonal level is to evaluate collections, information services, and technologies, to ensure that they are designed to facilitate effective interaction within and across cultures, including our own as well as the culture of users and creators.

2.2.1 Building cultural appreciation

Especially in situations where a professional exercises authority—that is, where an imbalance of power exists between practitioner and client—the outcome of an interaction may be affected substantially by the degree of cultural appreciation shown by the professional. In other words, it's not enough for the professional to understand the cultural differences that may be present in the interaction; rather, the professional needs to offer positive acknowledgement, approval, and respect for the client's cultural background and values. In this sense, cultural appreciation on the part of practitioners involves deliberately creating opportunities for clients—and by extension the communities they represent—to express their own ideas and expectations regarding the services they are to be provided.

2.2.2 Ethic of caring

The concept of “caring” has been developed in helping professions like social work and nursing to emphasize the importance of interpersonal relationships in professional work. A caring relationship is marked by authenticity—demonstrating in concrete ways that one is concerned with the outcome of a social interaction. Authentic caring requires reciprocity, building mutually beneficial relationships that

transcend mere obligation. For libraries, it's important to understand the value of users' affective or emotional needs as well as their intellectual need for relevant information.

As information professionals, we have a clear need to build an ethic of caring into the design of access systems and digital collections. One practical way to do this is to invest in continual, visible improvements to systems, and to closely monitor user interactions with systems to gather evidence on how they might be improved in the future.

2.2.3 Personal and cultural interaction

Ultimately, the process of acquiring cultural competence requires practice, through active engagement with users from diverse cultural backgrounds. In general, face-to-face communication affords a richer level of interaction than normally found online, as physical proximity and real-time interaction enable us to receive information in nonverbal as well as verbal form. Online communication may offer new opportunities for professionals to interact with creators and users of collections, but effective communication online requires close attention to the cultural barriers separating these communities.

2.2.4 Reflecting on values

As in the cognitive dimension, the process of building cultural competence requires us to regularly evaluate our interactions with users and creators, enhancing our cultural knowledge and enabling us to show responsiveness to diverse individuals and communities. As we suggested earlier, the larger aim of cultural competence is to institutionalize a process of continuous improvement in our information systems and services.

2.3 Environmental dimension

In the helping professions context, building cultural competence requires attention to a wide range of physical, geographical and societal factors such as language that shape cultures—beyond the set of values and interests we can attribute directly to a given community. Environmental factors also broadly affect the information environment of libraries and archives, of course, requiring practitioners to work actively to mitigate barriers inhibiting the appropriate use of collections and services.

3. Applying the cultural competence framework

Our starting point for this project was a series of case studies of cultural heritage institutions in southern Arizona, designed to assess the current state of the art for digitization and online access to collections related to the diverse cultures of the state. Our initial questions centered around institutions' motivation to digitize, their appraisal of items to include in digital projects, and how digital projects were being organized and carried out, especially in relatively small institutions marked by constrained resources and possessing collections representing historically underserved or minority populations in the state. In selecting institutions to include in the study, we had the benefit of a number of efforts in recent decades to

survey archival collections in the region and to identify communities lacking adequate documentation.²² After consulting local archivists, we developed a sample of 25 culture heritage institutions in the Tucson area. Of these, seven institutions representing six institutions agreed to be interviewed.

In general, our initial findings have reinforced the view that digitization activities in these institutions have been shaped more by resource constraints and organizational routines (especially centering on physical artefacts) than by conscious efforts to pursue online documentation strategies, whether of a formal or informal nature. This is not to say that institutions lack aspirations in this area, especially given the example of AMP. Indeed, one institution we studied, the Arizona Historical Society (AHS), has made a number of digitized collections available through AMP. Not surprisingly, we found that the Society's motivation to digitize collections was largely driven by the need to increase public awareness of the Society's collections. AHS itself is a visible presence in the state, with branches in four locations and a collection that includes roughly 700,000 historical photographs. However, budget cuts in recent years have greatly limited its capacity to undertake digital projects. For instance, the Tempe branch was actually forced to close down from 2010-12. It reopened with a total of four employees, all of whom have to act as archival generalists, taking turns on the reference desk, and having multiple tasks including processing and preserving physical collections.

Thus, to undertake digital projects, AHS has little choice but to rely on volunteers and student interns, placing a heavy burden on permanent staff to ensure quality in imaging and metadata production. At one time, AHS had a staff member dedicated to working with archival photographs, including reproduction and digitization, which is one reason the repository actually possesses a scanner. Still, with past as well as current resource limitations, AHS has encountered significant challenges in pursuing even small-scale digitization efforts involving a few hundred images. In many instances, a lack of standard descriptive metadata and consistent file naming practices has hindered efforts to appraise items and organize scanning projects. Likewise, the Society hopes to set up a dedicated online repository, and yet at present it lacks the technology infrastructure to do so at present. For this reason, a resource such as AMP has value for institutions as a content management system as well as a hosting service for online exhibits.

In fact, given the dependence of many Arizona cultural heritage institutions on the repository infrastructure provided by AMP, we decided to make a closer examination of a sampling of AMP collections, focusing on the metadata provided at both the collection-level and the item-level. Our aim was to understand the meaning of "best practices" for digitization in the local context of institutions and collections in Arizona.

In general, we found AMP to be a rich source of evidence on institutions' decision making both regarding appraisal and description in producing online documentation for a wide range of topics related to Arizona's diverse cultural landscape. More work is needed to provide a comprehensive analysis, and, using qualitative methods, especially to refine the cultural competence framework as a basis for analysing the results of digital projects. And so, for this project, we decided to focus on a sample of collection items to highlight some of the specific issues raised by the cultural competence framework in evaluating the outcome of digitization projects.

²² Linda A. Whitaker and Melanie I. Sturgeon, "The Arizona Summit: Tough Times in a Tough Land," *Journal of Western Archives* 1, no. 1 (September 2010): 2-28.

Example: Three Indigenous Baskets

1. Arizona State Museum: “Tohono O’odham Woven Plaque with Maze Design”
<http://azmemory.azlibrary.gov/cdm/singleitem/collection/asmspicer/id/10/rec/51>

This image was contributed to AMP by the Arizona State Museum, as part of a collection of digitized photographs by anthropologist Rosamond Spicer beginning in the 1940s. The Museum included a Web page with contextual information about the collection. The stated rationale for digitizing these materials was that in taking photographs, Spicer aimed to document events and activity in everyday life among the Tohono O’odham, especially during a nine-month period in the 1940s, when Spicer lived on the reservation as part of a funded research project.²³ A number of the digitized photographs added to AMP suggest that Spicer interacted directly with tribal members, but it’s not made clear how the resulting images were influenced by these interactions.²⁴

In fact, the Spicer photographs arrived at the Museum with a minimum of contextual metadata. Consequently, the metadata for the sample artefact is necessarily sparse—with the creator’s identity and the date marked as unknown, for instance. As for the item itself, the metadata record simply indicates that it’s a “Tohono O’odham woven plaque with maze design.” With no reference to the figure at the entrance to the maze, the larger cultural significance of the meaning is not apparent as it’s exhibited in AMP.

2. Pueblo Grande Museum: “Close coiled basket (2002.01.1)”
<http://azmemory.azlibrary.gov/cdm/singleitem/collection/pgmbasket/id/46/rec/3>

The Pueblo Grande Museum has contributed three collections to AMP, documenting indigenous cultures near the museum’s home in Phoenix. One consists of Hohokam artefacts (most dating between 1000-1500 ad). A second represents an early 20th century collection of Maricopa pottery, while a third featuring baskets from the museum’s permanent collection. Most of these were made by members of two related though distinct indigenous cultures: the Tohono O’odham (based in southern Arizona, along on the Mexico border) or the Akimel O’odham (based in central Arizona near Phoenix).

The basket in this example is labeled as a “man-in-the-maze” pattern, which the metadata identifies as a “common design for O’odham basketry.” On the meaning of the figure, the description indicates that the “maze is said to represent the house of Elder Brother, and symbolizes a person’s journey through life and search for balance.”

From a cultural competence perspective, we might expect this description to raise significant questions for users, especially calling for additional context about the Elder Brother figure and the story behind the maze. In this instance, the user is provided an indirect path to additional information, by way of the Salt River Pima-Maricopa Indian Community, which donated the basket to the museum, and, as the description explains, the Community uses man-in-the-maze image as its logo. No direct link is provided by AMP, but by navigating to the Community’s website, we can access a fuller account of Elder Brother and the maze. However, the image is not exclusive to the two tribes represented by the Community (Pima/Akimel

²³ <http://azmemory.azlibrary.gov/cdm/landingpage/collection/asmspicer>

²⁴ See, for instance: <http://azmemory.azlibrary.gov/cdm/singleitem/collection/asmspicer/id/54/rec/9>.

O’odham and the Maricopa/ Xalychidom Piipaash) and accounts of Elder Brother might be expected to vary, especially as the story has traditionally been transmitted orally. More importantly, given the spiritual nature of the image, we might argue that the description ought to include specific, authoritative information on the religious implications of the image and also the cultural context in which the basket was created and used.

3. Arizona State Museum: “Collection of Tohono O’odham Woven Containers”
<http://azmemory.azlibrary.gov/cdm/compoundobject/collection/asmtewes/id/66/rec/3>

This image is part of a series entitled: “Coiled Basket Making,” which is part of a larger collection contributed by the Arizona State Museum to AMP, consisting of photographs by photographer Helga Teiwes in the 1970s and 80s. In documenting the process of basket making, this series provides more detail than the Spicer collection in documenting the context of activities as well as artefacts produced by the Tohono O’odham. However, the image of finished baskets (shown grouped together) offers scant contextual information about the designs on the baskets. The description indicates that the baskets have “traditional and modern designs. Some of the designs are man in the maze, squash blossom, star and wheat.” Unfortunately, the metadata does not provide specific information about each item in the photograph, and it does not provide contextual information on the meaning or cultural significance of the designs, including the man-in-the-maze basket.

In spite of these shortcomings, the Teiwes collection (as depicted in AMP) shows a relatively high level of cultural awareness and personal engagement with the people represented in the photographs, especially by comparison to an early-20th century collection of Southwestern outdoor photography included in AMP by the Arizona State Museum: photographs by Forman Hanna, who was active in documenting the Arizona landscape as well as both the indigenous and American settler cultures.

Example: Arizona State Museum: Hopi images from the Forman Hanna collection

1. Hopi woman placing fuel around pots to be fired
<http://azmemory.azlibrary.gov/cdm/singleitem/collection/asmhanna/id/8/rec/16>
2. “Steps, Shipolovi”
<http://azmemory.azlibrary.gov/cdm/singleitem/collection/asmhanna/id/12/rec/19>
3. “Hopi girl sitting at the edge of mesa”
<http://azmemory.azlibrary.gov/cdm/singleitem/collection/asmhanna/id/13/rec/20>
4. “Girl dressed in traditional manta etc., standing near edge of cliff”
<http://azmemory.azlibrary.gov/cdm/singleitem/collection/asmhanna/id/14/rec/21>

From a cultural competence perspective, one factor that stands out in these images is the apparent cultural distance between Forman and his subjects. In the first example, we can infer that Forman intended to document an everyday activity in a naturalistic way (hence the three subjects are all looking away from the camera) and yet it’s not clear whether the photograph was staged—with the two children posed behind the woman in the foreground—or whether the subjects were even aware they were being photographed.

A similar naturalism (on Forman's part) can be found in the second photograph, which depicts a Hopi man at a distance from the camera as he descends stone steps built into a mesa. From the image itself and the accompanying metadata, it's not clear under what conditions the photograph was taken. Did the man agree to be photographed? Was he asked by Forman to hesitate near the bottom of the steps, an optimal position within the frame? As with the first image of the Hopi woman, did the man consent to be photographed? Such questions also apply to the third example, which shows a Hopi woman sitting on the edge of a cliff, at a distance from the camera. Unlike the photograph depicting pottery making, this image projects a distinct exoticism, both in the landscape and in woman's clothing and hairstyle. The accompanying image—the fourth example above—provides some context by showing the woman close up in traditional garb, and almost certainly posing for the camera. But here again, the interaction between photographer and subject is obscured by a lack of information. Did the subject make a special effort to dress up for the photographer, or did Hopi women at this time normally wear this type of clothing every day? Also, from all of the photographs, can we infer that Hopi culture was welcoming or at least neutral toward an American photographer?

Taken together, these examples highlight a number of theoretical and practical issues archivists face in pursuing online documentation strategies. Clearly, the AMP collections are relatively informal and limited in scope by the standard originally envisioned by Helen Samuels. Still, a “micro-scale” digitization effort such as the Helga Teiwes collection shows the potential for larger-scale digitization projects designed to aggregate digital objects and to add contextual information to existing collections available on the Web. We see the cultural competence framework as a potentially valuable tool for such efforts, especially in guiding the decision making process behind appraisal and re-appraisal of digital objects, and in preparing metadata.

One advantage in viewing digitization through a cultural competence lens is that it calls for transparency and self-examination by archivists as an integral part of the digitization process. For instance, in the AMP examples, some explanation of appraisal decisions was provided at the collection level, but the item-level metadata was opaque on the choice of items for inclusion in AMP. The sparsity of item-level metadata is not surprising given the general resource constraints on many first-generation digital projects, and the resulting technical limitations imposed by the use of Dublin Core elements as a default for online repositories. Using the cultural competence framework, archivists would begin digital projects with an assessment of their own culture—that is, activities, values, organizational structures, standard operating procedures in record keeping—while at the same time engaging in formal study of the cultures they are seeking to document.

In turn, the knowledge gained at the cognitive level should lay the groundwork for regular and productive interactions with members of the communities being documented.²⁵ In recent years, a growing number of archivists have called for closer and more effective collaboration between archivists and members of indigenous communities to ensure ethical treatment of artefacts, including repatriation and improved ways to describe the cultural context of items exhibited online.²⁶ Broadly speaking, we see the

²⁵ Bastian has explored in detail the need for deep engagement of local cultures in documentation efforts. Jeannette A. Bastian, “The Records of Memory, the Archives of Identity: Celebrations, Texts and Archival Sensibilities,” *Archival Science* (published online, 8 July 2012).

²⁶ Kimberly Christen, “Opening Archives: Respectful Repatriation,” *American Archivist* 74 (Spring/Summer 2011): 185-210.

cultural competence framework as supporting the emergence of best practices for managing digital cultural heritage, particularly as we seek to enhance the contextual information linked to digital objects on the Web. On this point, recent arguments on behalf of “participatory archives” and generally for archival practices that are “flexible, open, transparent, and collaborative”²⁷ also support our broader argument for cultural competence as a framework for digitization and the creation of online exhibits.

As noted above, the concept of cultural competence is rooted in the professional interactions that are fundamental to the human service fields. Of course as a professional culture, archivists have traditionally played a rather different kind role in society, especially as records and artefacts are a medium of communication across time and distance.²⁸ As many archival theorists have noted, the highly complex role archivists play in shaping historical and cultural understanding places a heavy burden on the profession, especially as we seek to define standards and best practices that balance practical constraints on recordkeeping against the need to support the highly diverse modes of understanding and interpretation that make up archives as a form of cultural heritage.²⁹

Ultimately, in applying the cultural competence framework to the problems we face in digitizing archives, it becomes clear that the existing environment—both in the physical legacy of paper records and in the technical affordances of the Web and online repository systems—calls for archivists to set explicit constraints—both positive and negative—on digitization to make online documentation a sustainable enterprise. By negative constraints, we mean that archivists have to take into account institutional capacity in appraising records—thereby limiting the scope of online collections. By positive constraints, we mean that for digital repositories to achieve cultural (as well as organizational) sustainability, archivists have to pursue explicit documentation strategies designed to capture—as much as possible—the full diversity of artefacts and cultural meanings implicit in archival collection.

²⁷ Kate Theimer, “What is the Meaning of Archives 2.0?” *American Archivist* 74, no. 1 (Spring/Summer 2011): 58-68, p. 68. See also, Isto Huvila, “Participatory Archive: Towards Decentralised Curation, Radical User Orientation, and Broader Contextualisation of Records Management,” *Archival Science* 8, no. 1 (March 2008): 15-36.

²⁸ On the ways in which archivists influence historical understanding, see: Joshua Sternfeld, “Archival Theory and Digital Historiography: Selection, Search, and Metadata as Archival Processes for Assessing Historical Contextualization,” *American Archivist* 74, no. 2 (Fall/Winter 2011): 544-575.

²⁹ See, for instance, Eric Ketelaar, “Cultivating Archives: Meanings and Identities,” *Archival Science* 12, no. 1 (March 2012): 19-33.

Preserving Digital Heritage

The UNESCO Charter and Developments in the Netherlands

Marco de Niet, Titia van der Werf and Vincent Wintermans

Abstract

The e-Depot of the Netherlands National Library and Images for the future, the mass-digitization project for audiovisual heritage from the Netherlands carried out by the Institute for Sound and Vision, are two large scale projects for long-term preservation and digitization in the Netherlands. Comparing the results achieved and the challenges met in these projects, the study gives insight in how assumptions and problems in the field of digital preservation have evolved in the past decennium. This will fuel the discussion on the revision of the Charter for the Preservation of the Digital Heritage.

Authors

Marco de Niet is the director of the DEN foundation, the Dutch knowledge centre for digital heritage. He studied History of the Book at Leiden University. He is actively involved in the Europeana Network and core partner in the ENUMERATE project that measures the progress of digitization in Europe. He is a founder member of the Dutch Memory of the World Committee, a board member of the Dutch Museum Register and a member of the Council for Dutch Language and Literature. Before DEN he worked at the National Library of the Netherlands as head of Innovative Projects and Digital Preservation.

Titia van der Werf is senior programme officer at OCLC Research, based at OCLC's office in Leiden. Between 1993 and 2002 she worked for the R&D department of the National Library of the Netherlands and contributed significantly to early research into digital preservation and design of the first e-Depot system. She was actively involved in the development of the Open Archival Information System (OAIS) Reference Model and was on the team that drafted the UNESCO Charter on the preservation of digital heritage (2003). She is a member of the Netherlands Memory of the World Committee.

Vincent Wintermans studied Slavic languages at Leiden University. He is policy officer at the Netherlands National Commission for UNESCO and an observer attached to the Netherlands National Committee for the Memory of the World Programme. His current professional interests are open access publishing, documentary heritage, endangered languages and communication & development. He carries out research on the history of the work of UNESCO and the IIC in the field of linguistics and bibliography.

1. Introduction

UNESCO's General Conference adopted the Charter for the Preservation of the Digital Heritage¹ in 2003. This non-binding standard-setting instrument, targeted mainly at governments, explains the particular vulnerability of digital heritage, asks for immediate action to protect it and recommends the development of strategies and international cooperation. In the same year, UNESCO published Guidelines for the Preservation of the Digital Heritage,² written by Mr Colin Webb of the National Library of

¹ Charter for the Preservation of the Digital Heritage: http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html.

² Guidelines for the Preservation of the Digital Heritage:
<http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>.

Australia. These Guidelines elaborate the issues raised by the Charter in great detail, both from a management perspective and from a technical and practical perspective.

Article 12 of the Charter requests UNESCO to ‘determine, on the basis of the experience gained over the next six years in implementing the present Charter and the Guidelines, the need for further standard-setting instruments for the promotion and preservation of the digital heritage’. The results of the survey conducted in 2009 were unsatisfactory: only a small number of Member States responded and the outcomes were diffuse. Some countries argued for a new standard-setting instrument, or for the upgrading of the Charter to a higher level. Some asked for efforts to make the Charter better known.

In 2009, the Netherlands National Commission started a project with the aim of answering on a more empirical basis the question of whether a revision of the Charter and the Guidelines is possible and necessary. In a workshop organised in the framework of the IV International Conference on the Memory of the World Programme *Culture-Memory-Identities* (Warsaw, 18-21 May 2011), it presented and analysed examples of how Charter and Guidelines have been used in South Africa, Brazil and Poland. The results of this preliminary step, to be published in the proceedings of the conference,³ show that the Charter has been a useful instrument in the three countries considered. Some themes that might be included in an addendum to the Guidelines were tentatively proposed: technical innovations like cloud computing, the appearance of commercial actors in the field of digital preservation, the experiences of non-Western countries in digitising their heritage, and the responsibility for digital heritage that cannot be attributed to a country.

When UNESCO announced that the revision of the Charter would be one of the subjects to be treated at the Conference *The Memory of the World in the Digital Age*, the Dutch Commission decided to analyse experiences gained in the Netherlands with long-term mass preservation of digital heritage in order to detect major developments in the field that could be the subject of further standard-setting by the Organisation.

2. Digital preservation in the Netherlands

Dutch heritage institutions participated in the very first European projects to develop methods and technologies for digital preservation in the 1990s. Soon after, the Dutch government was actively involved in the realisation of the UNESCO Charter on the Preservation of Digital Heritage. The National Library of the Netherlands (Koninklijke Bibliotheek, KB)⁴ was the first national library in the world to present an operational longterm archives for electronic journals, in part thanks to a substantial grant from the Dutch government. With a similar governmental commitment, the Netherlands Institute for Sound and Vision (Nederlands Instituut voor Beeld & Geluid, BenG)⁵ became one of the international front-runners of long-term preservation and management of digital audiovisual heritage.

In 2007, the KB and DANS, the national scientific data archive,⁶ took the initiative towards a national and more structured approach to digital preservation of public sector information. Following the

³ V. Wintermans, “An Addendum to the Charter and Guidelines for the Preservation of the Digital Heritage,” http://www.unesco.nl/documents/documenten-natcom/Addendum_Preservation_Digital_Heritage.pdf.

⁴ www.kb.nl

⁵ www.beeldengeluid.nl

⁶ www.dans.knaw.nl/

examples in the United Kingdom (Digital Preservation Coalition)⁷ and Germany (NESTOR),⁸ both organisations invited other public institutions with digital preservation as a core task to join forces. Together they founded the National Coalition on Digital Preservation (NCDD).⁹ The mission of this foundation is “to ensure that there will be a stable organisational and technical infrastructure in the Netherlands for the preservation of and permanent access to digital information that is relevant to science, culture and society at large.”

The first assignment for the NCDD was a national scan of the digital preservation situation in the sciences, the public media, the cultural heritage sector and the government. The results were published in the rapport *A Future for our Digital Memory* (July 2009).¹⁰ From the scan, it was obvious that the role of some front-runners was crucial for the progress made in the various sectors: the e-deposits of DANS and the KB took care of scientific data and publications, respectively; the National Archives and some innovative archives, including the city archives of Amsterdam and Rotterdam, laid the foundation for a networked archiving system for governmental documents and data. BenG built the largest public digital archive in the country for audiovisual materials. The weakest sector was the cultural heritage sector, notably the museums. In this sector, there was not one obvious front-runner to take up the responsibility for developing and managing a digital archive for cultural data and objects. Following on from this conclusion, various cultural institutions took the initiative to install a Cultural Coalition on Digital Preservation. However, this Coalition currently limits itself to raising awareness, developing policy and sharing knowledge and research, given its lack of resources and the lack of an organisational infrastructure to set up such a cultural digital archive.

The time when digital preservation was mainly a research topic is well behind us. Digital information is omnipresent and the lack of professional digital preservation strategies and solutions will inevitably result in loss of public data. In the past decade, large-scale systems for the preservation and management of digital data have become operational in the Netherlands. However, these infrastructures are not yet fully mature. The archiving systems are constantly under development and are facing a rapidly growing flood of digital information that seems to be in a permanent state of flux. The institutions that manage these systems are still trying to find reliable financing and business models to sustain the high costs of continuous development and maintenance into the future. Also, the digital collections of many institutions are not yet involved in digital preservation solutions. With the current economic downturn and a government that is cutting its budgets for science, culture and the public sector as a whole, it is a big challenge to take digital preservation to the next level. This article is intended to provide some inspiration for future developments by presenting the lessons learned from two Dutch pioneers that have built large scale systems for the preservation of digital objects that fall within the scope of UNESCO’s Charter on the Preservation of Digital Heritage.

3. Introduction to Beeld en Geluid (BenG) and the Koninklijke Bibliotheek (KB)

The Nederlands Instituut voor Beeld en Geluid (Netherlands Institute for Sound and Vision, BenG) is a cultural-historical organisation that collects, preserves and opens up the Dutch audiovisual heritage for

⁷ www.dpconline.org

⁸ www.langzeitarchivierung.de

⁹ www.ncdd.nl

¹⁰ www.ncdd.nl/en/publicaties.php

various user groups, such as media professionals, education, science and the general public. It is one of the largest audiovisual archives in Europe. It is located in the city of Hilversum and is housed in one of the most eye-catching buildings in the country. The collection contains more than 800,000 hours of television, radio, music and film from the beginning in 1898 until the present day, including the complete radio and television archives of the Dutch public broadcasters. The institute manages over 70 percent of the Dutch audiovisual heritage. Digitization is a core task of the institution, both for conservation and for enhancing access.

The Koninklijke Bibliotheek, National Library of the Netherlands (KB), was founded in 1798 and is based in The Hague. The KB is the one of the largest libraries in the Netherlands together with the university libraries of Amsterdam and Leiden. The KB is both a research library and, since 1974, a deposit library. The Netherlands is one of the few countries in Europe where there is no deposit law for books and publications. The deposit library is based on mutual agreements with publishers and their representatives. The KB became an independent body in 1982. In the early 1990's, the KB set up its own research department, which helped the library to modernise its services and processes and acquire a strong international profile. The current policy plan of the KB puts digital information at the heart of its mission.

4. Mass digitization as a catalyst

In the past decades, major infrastructural developments and projects in the Netherlands were financed from the profits made from exploiting the country's gas reserves. The Dutch government is closely involved in this industrial enterprise, and the Dutch treasury benefits each year from several billions of euros. Until recently, special governmental programmes were set up to invest these extensive tax revenues in all kinds of public services, including digital infrastructures for science and culture. Science and culture are, together with education, the main areas of the Dutch Ministry of Education, Culture and Science (Onderwijs, Cultuur & Wetenschap, OCW).¹¹ Both institutions described in this paper are supported and financed by this Ministry. BenG is organisationally positioned within the directorate for Media, Literature and Libraries, while the KB is within the directorate of Research & Science. Both directorates have invested extensive funds in digitization programmes and digital preservation systems to support the digital transition in these institutions.

In 2000, more or less simultaneously with their project to develop an e-Depot for electronic publications, the KB initiated the Memory of the Netherlands programme for the digitization of cultural heritage. This programme, inspired by the American Memory project of the Library of Congress,¹² consisted of three components: 1) a funding scheme for smaller institutions to have visually interesting collections digitized, 2) an expert centre to set up quality guidelines for digitization of cultural heritage, and 3) a website through which all the digitized collections could be accessed.¹³ Currently, the website contains close to 800,000 cultural objects from 100 institutions. It was intended to have the digital masters created in the Memory of the Netherlands programme stored in the e-Depot of the KB. This plan has not yet materialised due to technical constraints of the e-Depot system, which was primarily designed for the preservation of scientific journals, but also due to the lack of a proper business model to sustain and preserve these digitized collections over a longer period of time (see also below).

¹¹ www.rijksoverheid.nl

¹² <http://memory.loc.gov/>

¹³ www.geheugenvannederland.nl

In July 2007, an even larger project was launched for the restoration, preservation and digitization of Dutch audiovisual collections, *Beelden voor de toekomst* (Images for the Future).¹⁴ This project was initiated by BenG and the digitization included the collections from the Dutch Film Museum EYE and the photographic collections of the National Archives. The total budget was originally 154 million euro, of which 64 million had to be paid back to the government before the year 2025, e.g., from revenues earned through cultural entrepreneurship with the digitized collections. In 2010, a mid-term evaluation by the Dutch independent research organisation TNO made clear that these expectations could not be met, and the overall budget was reduced to 115 million euro. This means that less material can be preserved and digitized, but the obligation to repay some of the funds was dropped altogether. Currently, almost 85,000 hours of video, 84,000 hours of audio, 16,000 hours of film and close to 2.5 million photos have been digitized. As for BenG, their budget within the project is 85 million euro and a total of 200,000 hours of material will be preserved and digitized.

For both the KB and BenG, these large-scale digitization projects influenced the setup of the digital preservation systems, as these systems had to support the ingest of both new (born digital) content from external suppliers as part of their ongoing mission (publishers and broadcasters respectively) and of digitized counterparts of historic collections of various origins. In terms of the UNESCO Charter on the Preservation of Digital Heritage, both institutions were in a good starting position to address the major challenge of digital continuity in all its complexity (Charter, Article 5).

5. System development and technology partners

In 2000, the KB and IBM started building the Digital Information Archiving System (DIAS) that was to become the technical core of the infrastructure for KB's e-Depot for electronic publications. It was clear from the start that this project could not rely on out-of-the-box solutions alone, because at that time no solution readily addressed both the aspects of large volume and durable storage as well as the long-term preservation requirements. As well as implementing the e-Depot system, IBM, KB and the British Library carried out a "Long-Term Preservation Study"¹⁵ to define the requirements of the OAIS Preservation Planning function as an extension of the basic content management and storage functionality provided by the initial DIAS version. By 2012, the KB had decided that DIAS was no longer a viable solution and terminated its contract with IBM. What had happened?

There were two main reasons for KB's decision. As the KB needed to ingest more and more different types of content, coming from other sources than e-publishing (specifically digitization and web archiving), the operational DIAS system did not evolve sufficiently, and the ingest procedure became a bottleneck. The system did not support important standards or interoperable interfaces to connect with preservation tools and methods that were coming out of research projects. The negotiation process with IBM for requesting changes and new functionality and getting rapid software upgrades was not without its frictions. KB experienced this as a vendor lock-in constraint.

It was also a setback that DIAS did not evolve into a widely used market product, which was the original intent and hope of both parties. With only the KB in the Netherlands and the Deutsche Nationalbibliothek (DNB) in Germany as DIAS users, the product could never become profitable for IBM

¹⁴ <http://beeldenvoordetoekomst.nl/>

¹⁵ www.kb.nl/hrd/dd/dd_onderzoek/reports/1-overview.pdf

and the costs remained high for the customer libraries. As a result of KB's decision to terminate its relationship with IBM, the e-Depot is currently undergoing the most fundamental migration in its history.

DIAS is being replaced by a totally new system that is designed, built and implemented in-house. The KB has come to the conclusion that there are no adequate systems for its e-Depot in the marketplace and that they do not want to repeat the vendor lock-in experience. Open-source products are integrated in the new system, if fitting the requirements—for example, software to support some of the workflows. The digital preservation approach implemented in DIAS with the “Reference Platform” infrastructure and view paths for specific digital object types is done away with. The new system is specified on the basis of 15 years of R&D into digital preservation and more than 10 years of practical experience with the e-Depot service. The whole operation includes a medium migration exercise—migrating from optical disks to hard disks—and a complete redesign of the metadata management—migrating the storage, retrieval, discovery and preservation metadata into a single system. The archival information packages (AIPs) will be repackaged from a tar file into a PREMIS-XML file. The KB expects that adopting the PREMIS standard will allow for a better administration of the ‘events’ that objects in the repository undergo—which is crucial information for the long-term preservation, as migrations and other transformations will continue into the future. As issues of scalability and rapid change will not disappear, the KB is preparing itself for recurring system migrations in cycles of 10 to 15 years.

The digital archives of BenG had completely different beginnings, but there are interesting similarities to note in the further developments. The creation of the digital archives in Hilversum was not a conscious decision of a heritage institution to start expanding its usual business into the digital domain, but the inescapable result of the decision taken by the Dutch public broadcasters to digitize the production process. By 2006, the whole broadcasting production process had been digitized.

For the technical infrastructure of its archives, BenG had always relied on Nederlandse Omroepproductie Bedrijf (NOB), an audiovisual production company that was part of the Dutch public broadcasting landscape, and a close neighbour of BenG in more than just the geographical sense. Also in 2006, NOB Cross Media Facilities was taken over by Technicolor, a division of the French multinational Thomson. Both changes (digitization and privatisation) evidently had major effects on operations at BenG. Thomson set up the archives as a storage-as-a-service facility, meaning that BenG paid per unit of storage capacity used. With the increasing volumes of digitized content coming out of the production process and the Images for the Future digitization project, costs rose at an alarming rate from 2007 onwards. BenG tried to disentangle itself from Thomson, but this proved to be a difficult operation because the storage-as-a-service model meant that Thomson was providing services to other clients using the same infrastructure. At the time of its conception, this “shared platform” philosophy was thought to be a cost-efficient solution. Over time, BenG found itself to be the largest customer by far, but being constrained by obligations from Thomson to other, smaller clients. BenG experienced this as a vendor lock-in constraint. Because of the many dependencies, it concluded that tendering for a new archiving solution was impossible and decided to build a completely new archiving facility from scratch.

This was a major decision because it entailed that a vast body of technical knowledge had to be acquired by its staff in order to catch up on archiving expertise and skills that had previously been outsourced to Thomson. Disentanglement of the processes, requirement specifications, design, development and implementation of a new archiving environment all had to be carried out in parallel with the continued and uninterrupted service delivery to the broadcasters. According to BenG, its new environment operates at far lower costs.

The in-house system of BenG is built to support critical stages in the heavy-weight broadcasting production process: each production is immediately archived and available for reuse after broadcast in the controlled environment for the Hilversum broadcasting companies. Storage requirements are evidently geared to high-volume storage, as the archives have now reached 5PB of capacity usage, and the high-availability requirements do not allow the use of cheaper and slow access media, such as tape storage. Preservation functionality was not a high-priority for BenG when they started their in-house development project. Medium migration has been the obvious preservation action in 2011 and 2012 and is not considered to be very complex. Formats seem to have a longer lifespan than expected a few years ago, so format migration is not an immediate concern.

When asked about the lessons learned, both KB and BenG give the same answer: do not buy solutions or outsource tasks that you do not fully understand. Make sure you have the expertise in house to specify everything you need and to enter into dialogue with vendors and technology partners at a professional level.

Although the term ‘vendor lock-in’ does not occur in the Charter or Guidelines for the Preservation of the Digital Heritage, both documents have something to say about cooperation with industries. The Charter stresses the need for reliable systems (Article 5) and the need to share technical knowledge and experience between industries and heritage institutions (Article 11), but it does not warn against overdependence on parties that, by their nature, are primarily driven by profit. The Guidelines (p. 113) point to the benefits of purchasing off-the-shelf solutions, such as storage, which is a core area to digital archiving, but do not provide guidance in forging sustained business relationships with parties in the IT industry. Such relationships require an understanding of and respect for each other’s interests, in particular in new areas where market opportunities still need exploring and requirements are still developing. Digital preservation is clearly such an area: both the industry and the community of practice have not yet fully matured. Defining generic requirements and designing robust systems for the market requires close collaboration between the community of practice and the industry.

Interestingly, both the KB and BenG have “disentangled” themselves from their industry partner and they have opted for in-house development and as much use of open standards as possible.

Open-source software is incorporated where useful. Open-source development is not a priority for KB and BenG, considering the specialised nature of some of the requirements and the limited number of potential members of that community of practice. Being neither open source nor vendor solutions, the development and maintenance costs of the in-house solutions of BenG and KB will therefore remain high.

6. Geared to growth

The KB and BenG are both transitioning to large-scale archiving systems that can accommodate all their digital collections (both born-digital and digitized). The KB no longer calls its system e-Depot, but refers to it as their Digital Stacks (‘Digitaal Magazijn’) to underscore the multi-purpose aspect of the archive. In the past 15 years, both institutions have not only witnessed a remarkable growth in scale, but they have also had to deal with a growing variety of content. This leads us to the question: how did their collection policies and selection criteria evolve?

As an institution, BenG is the product of the merger in 1997 between the business archives of the public broadcasting service, the film archive of the Netherlands Government Information Service, the Film and Science Foundation (Stichting Film en Wetenschap), and the Dutch Broadcasting Museum (Nederlands Omroepmuseum). The core and largest collection, however, is the material produced by the

Dutch public broadcasting service. No initial selection of born-digital broadcast materials takes place; all are automatically ingested in the system on a daily basis. This has been done at the request of the broadcasters. In the longer run, however, the Institute expects to select, just as it did in analogue times. It will then, for example, store the entire series of news bulletins, but only a few instalments of popular daily game shows.

BenG's policies are geared to growing its mass archiving facility and allowing AV collections from other institutions to make use of it. In other words, the institute is not looking for ways to manage growth by selecting up-front. On the contrary, it is stimulating growth in search for economies of scale. Likewise, the objective of the Images for the Future digitization programme is to digitize as much as possible without applying selection criteria. Selection is very time-consuming, it argues, with most of the project budget for digitization being lost on this activity. Moreover, selection has already taken place during the analogue collection building process and should not be repeated. Not all collections can be digitized, however, so selection at the collection level does still take place. In the bewildering mass of different file formats, and confronted with legacy metadata of hugely varying quality, the institute looks for 'sweet spots', an optimal balance between production volume, available budget, time constraints, quality and present and future archival, preservation, access and repurposing requirements.

At the end of the 1990s, the KB had extended its voluntary deposit agreement with the Dutch Publishers Association (Nederlands Uitgeversverbond) to cover electronic publications (offline and online). In 2002, a bilateral e-archiving agreement was signed with Elsevier and included all Elsevier journal titles. After these major achievements and the successful implementation of its e-Depot system at the end of 2002, the KB felt confident enough to open up its facilities to other major international journal publishers.¹⁶ Currently it has a storage capacity of 12Tb and contains over 18 million digital publications (mostly scientific articles). Its objective is to collect 80% of the output of the STM publishers.

Alongside the international e-Depot, the KB maintains another digital archive for the digitized collections that are the result of KB's mass digitization programmes: Metamorfoze,¹⁷ Memory of the Netherlands,¹⁸ Newspapers Online¹⁹ and more recently, the Early Dutch Books Online programme, which aims to digitize all Dutch titles published between 1750 and 1940 held at the KB and the university libraries of Leiden, Amsterdam and Utrecht.²⁰ The books digitized under the Google contract are also part of this programme. The digitized collections currently total a storage capacity of 470 Tb. The KB has also been archiving a growing selection of Dutch websites since 2005 and it keeps back-up copies of the Dutch university institutional repositories. All these digital collections have not been ingested into the current e-Depot, but will be incorporated in the new Digital Stacks.

Just like BenG, the KB does not use collection policies and selection criteria as means to control growth. The collection and acquisition policies of both institutions are geared to growth. Both are investing in large-scale digital back-end infrastructures and facilities in order to be able to offer these to other parties (e.g., publishers and memory organisations) with a view to achieving economies of scale and recovering costs. Interestingly, the KB has started to develop a valuation model which can help in judging the value of a collection; different preservation levels can then be offered at different costs based on this.

¹⁶ <http://liber.library.uu.nl/index.php/lq/article/view/7866/8062>

¹⁷ www.metamorfoze.nl

¹⁸ www.geheugenvannederland.nl/?en/homepage

¹⁹ <http://kranten.kb.nl/>

²⁰ www.earlydutchbooksonline.nl/en/edbo

The Charter declares that ‘the main criteria for deciding what digital materials to keep would be their significance and lasting cultural, scientific, evidential and other value’ (Article 7). The KB example shows that these criteria can also play a role in deciding on the preservation regime to be applied to a specific collection.

7. Publishing and archiving formats

A selection criterion that has been discussed in the digital preservation debate since the late 1990s is the format of the content. It has been argued that memory institutions with a mission to preserve digital materials should promote best practices with the creators of e-content, and discourage the use of formats that are more prone to obsolescence than others. Some archives have actually followed this principle by limiting the variety of formats in which content files could be submitted. What is the experience of BenG and the KB?

The e-Depot of the KB generally receives PDF files, but there is a clear trend of diversification of incoming formats, as scientific publications are increasingly accompanied with data sets and other materials (‘enhanced publications’). Moreover, as we saw above, the KB itself is increasingly using its digital stacks for other materials than publications, which leads to an even larger variety of formats to be ingested in the e-Depot.

The BenG archives receive formats that are generally created by trusted parties, like the MXF format supplied by the public broadcasters. In most cases, BenG has control over the formats, in particular the digitized products created in house. It has just started the largest film digitization project in the world with the objective to preserve deteriorating analogue film materials. It uses the DPX standard, which is less suitable for direct online access. Professional users will get previews and the possibility to order a copy. Consumers will have to pay a relatively high price for access to the DPX films.

Both organisations promote best practices and publish guidelines about the best formats for archiving, but they do not restrict submissions of suppliers on the basis of formats. An important argument for adhering to these guidelines is that this enables the archiving institution to guarantee long-term access.

The Charter states in Article 5 that ‘long-term preservation of digital heritage begins with the design of reliable systems and procedures which will produce authentic and stable digital objects’, and in Article 11 that ‘industries, publishers and mass communication media are urged to promote and share knowledge and technical expertise’.

The Guidelines stress that long-term preservation is only possible when preservation issues are taken into account from the very beginning, i.e., when a digital item is created. It is critical about the role of the industry: ‘currently, preservation efforts have to work against the prevailing trend of digital technology, and how it is developed and used’. (5.2.5) Again, in chapter 13: *‘digital materials are created by producers who are not necessarily concerned with long-term availability. [...] Without some kind of intervention, it is unlikely that digital heritage materials will automatically be made in ways that will minimise costs and remove barriers to preservation’*. (13.4) More specifically about file formats, the Guidelines remark that creators ‘should be encouraged to use very widely adopted, well-standardised file formats that fit their purposes. Generally speaking, data in simpler formats using open source, non-proprietary software are easier to preserve (although some proprietary applications achieve such widespread use that they may be accepted as an industry standard, especially if their specifications are openly published). [...] If access or copying barriers are considered necessary to protect intellectual

property, they may well make preservation impossible. Arrangements will be needed to allow preservation processes such as copying to take place'. (13.13)

The ten years of experience at the KB and at BenG do not seem to support the worries of both Charter and Guidelines concerning the file creation format. The managers of the two archiving services have built relationships of trust with the publishing and broadcasting industry, respectively, and they are confident that these parties share an interest in the long-term availability of their content.

8. Finances and business models

“The costs of preservation programmes are hard to estimate because they encompass so much uncertainty, including evolving techniques, changing technologies and very long timeframes. [...] Total costs are also likely to remain high, including set-up costs and significant recurrent costs.” (Guidelines, p. 23).

Based on the experiences in the Netherlands, it is safe to say that the costs of development and operational management of large-scale digital repositories, such as those at BenG and the KB, can easily run into millions of euro per year.

Both systems started as project-based preservation programmes, with support from the national government in an international (mostly European) research context. From the beginning it was clear that the goal had to be the establishment of large-scale operational systems. Both institutions were confronted with a fast growing influx of digital materials that had to be dealt with as part of their mission. This probably made it easier for them to present compelling cases of urgency to the Dutch government to invest in their e-deposits. The KB got a fixed addition to their annual budget for running the e-deposit of international publications. It was calculated that the current annual cost of the international e-deposit is around 1.3 million euro per year. It is obvious that costs will rise considerably once the digitized collections that were described above have been added.

BenG receives a budget of 2 million euro per year to maintain its digital archive. However, with a storage volume of an impressive 5 Pb (growing to 10 Pb by the end of 2014), this will not prove to be sufficient. It is estimated that an annual budget of over 3 million euro would be needed to maintain the system and services that come with it.

Besides governmental funding, both organisations have committed a lot of their own resources as well, for instance by organising temporary matching budgets to (research) grants and by permanently reallocating budgets to new tasks. With the maintenance of trusted digital repositories, there are recurrent costs “associated with staff, accommodation, energy supplies, network use, telecommunications costs, storage media such as disks and tapes, and consumables” (Guidelines, p. 56). With linear or maybe even exponential growth of the content of digital archives, these cost factors can easily run out of control.

“Reliable preservation programmes must be sustained over long periods, so they require business models that guarantee adequate resources will continue to be available. Unfortunately, such guarantees are rare in the real world. Most programmes have to survive with less certainty.” (Guidelines, p. 54).

The European Commission and the Dutch government support the emergence of more public-private partnerships (PPP) in the area of digitization and digital infrastructures in order to enlarge private investments. As described, BenG and the KB both engaged in public-private partnerships. In general, it can be stated that these PPPs were quite successful during the research and development stages, as both parties were willing to invest money and knowledge. For the operational management, this relationship turned into a more traditional supplier-client-model. In that respect, it should be remembered that “while suitable service providers may be found to carry out some functions, ultimately responsibility for

achieving preservation objectives rests with preservation programmes, and with those who oversee and resource them” (Guidelines, p. 24). The Guidelines clearly state here that the ultimate responsibility for digital preservation cannot (or should not) be privatised. BenG and the KB have both secured their responsibilities and they have tightened, or in the case of the KB completely taken over, control of the development and maintenance process. The Dutch cases seem to indicate that it is probably easier to set up PPPs around research and development than identify business opportunities with the operational management of digital archives. In this respect, the concern should be noted that the European Commission does not address digital preservation as a separate research topic anymore in their Horizon 2020 strategy. This may seriously reduce the opportunities for continued PPPs in digital preservation.

Besides public-private partnerships as an additional source of funding, there is also another opportunity: exploiting the facilities of the large-scale digital archives by providing digital preservation services to smaller affiliated institutions. Both BenG and the KB have experimented with this, but without much success. The KB tried to offer its repository as a facility for storing the master files of digitized cultural heritage resources from other institutions. The KB did a feasibility study to set up such a ‘TIFF archive’ for smaller archives, libraries and museums, and the conclusion was that (for now) a business model that was acceptable to all parties involved could not be achieved. It turned out that most institutions were not willing to pay (yet) for storing the master files in a central facility. BenG ran a similar project called ProArchive, which also failed to find a suitable business model. One of the lessons learned from the ProArchive project was that services for permanent storage should be separated from services for metadata management. Because of the combination, many institutions were not interested in participating, as they required different metadata management options than those that BenG provided.

Another possibility for creating additional revenues is charging end users for content delivery. This is complicated for a variety of reasons. KB and BenG are both non-profit organisations, and cannot participate in economic competition as private partners can. In addition, they do not own many of the rights attached to the content in their archives. It would take a lot of resources to clear those rights, so it is not clear from the outset what the economic benefit for the institution itself would be.

A final possibility for sustaining the digital archives is to implement a mechanism of payment upfront, rather than charging the user: the party that delivers the content pays for the cost of digital preservation. The main motivation for such an approach would be to strengthen the joint responsibility for continuity of the preservation value chain. Publishers, scholars and archiving institutions, all have their own interests in achieving that continuity. Publishers can use the public digital archives for keeping their content alive; archiving institutions can get rid of complex licensing agreements, which may jeopardise long-term access once an agreement has expired. Scholars have an interest in safeguarding the integrity of the scientific record.

Services like Portico²¹ show that collaboration between libraries and publishers can be beneficial to both. The KB is exploring the options to provide services in Europe that are similar to those of Portico. It aims to run the international e-deposit of scientific journals without being dependent on additional government funding anymore.

²¹ www.portico.org

9. Organisational impact

Although some of the traditional tasks of cultural heritage institutions provide a good foundation to accept digital preservation responsibility (Guidelines p. 49), it is obvious that digital preservation programmes also require new skills and expertise and bring new ways of working, maybe even new organisational structures to cultural heritage institutions.

In recent years, there have been various initiatives in the Netherlands to better understand the digital transition that is currently taking place in cultural heritage institutions. Both quantitative and qualitative research among archives, libraries and museums have shown that there is hardly any activity left that has not been affected by the emergence of ICT.²² The need to share knowledge, experiences and best practices has been a strong driver to organise various national and international conferences, workshops and other meetings on digital preservation issues, including file formats, standardisation, business models, system development, acquisition policies, emulation etc.

Because of their long-standing involvement with digital preservation, both the KB and BenG are living proof of the impact of ICT on the mission and main tasks of cultural heritage institutions. In the past decade, both institutions have become international expert centres on digital preservation. The KB is a founding member of both the Alliance for Permanent Access (APA)²³ and the Open Planets Foundation (OPF),²⁴ a community hub for digital preservation whose main goal is to jointly manage and improve tools and research outcomes. BenG houses the European PrestoCentre, a membership-driven organisation that brings together a global community of stakeholders in audiovisual digitization and digital preservation to share, work and learn.²⁵ BenG also maintains an extensive knowledge base in Dutch on preservation and digitization of audiovisual collections.²⁶

How fundamental some of the changes in the ways of working in both institutions have been, can be illustrated by the Images for the Future project at BenG and by the Metamorfoze programme at the KB. Metamorfoze is the national programme for the preservation of paper collections (in libraries and archives) in the Netherlands. Until only a few years ago, both programmes considered preservation on film to be the best preservation strategy for fragile physical originals. More or less simultaneously, they radically changed the preservation strategies, abandoned the use of analogue copies and made the switch to preservation digitization and preservation imaging. As a consequence, all guidelines had to be redefined and expertise in this area had to be safeguarded. As Hans Westerhof, Vice-Director of BenG puts it: The new archivist is someone who understands networking and storage, he or she is a computer scientist or an information analyst.

As for their internal structures, both organisations have strong research departments which operate independently from the daily operations of the digital archives. Looking at the various organisational models described in the Guidelines (p. 59), the KB's model can probably best be described as "a series of specialist units looking after different aspects," while BenG's model is "a matrix of people working in different areas, responsible to an overall programme manager." Both institutions made the deliberate decision to invest in their own staff to increase their knowledge of digital preservation to become less

²² www.den.nl/art/uploads/files/Publicaties/BusModIn_eng_final.pdf

²³ www.alliancepermanentaccess.org/

²⁴ www.openplanetsfoundation.org/about

²⁵ www.prestocentre.org/about-us

²⁶ www.avarchivering.nl/

dependent on external companies and be more in control. As Westerhof puts it: “Don’t buy what you don’t understand.”

As rather large cultural institutions, they are able to make these kinds of investments in hiring experts as part of their regular staff. They both agree that smaller institutions should be able to tap into this new kind of knowledge and get their support. This is in accordance with the Guidelines, which states that not all institutions with a traditional heritage role “should try to become digital heritage managers: in some cases the resources and expertise required are just not available” (Guidelines, p. 49). However, this does not mean that the KB and BenG do not acknowledge that there is no additional expertise available at smaller institutions. There are several institutions who have specialised in the preservation of specific collections, such as born-digital art. The KB and BenG also recognise that other institutions have their own responsibilities regarding the preservation of their collections. Building an inclusive community of practice, however, proves difficult to realise; there are many factors that can get in the way of closer collaboration, such as lack of trust, no appropriate business models, lack of common standards, lack of proper agreements on selection and acquisition and the absence of open technology. The challenge for the entire heritage community in the Netherlands and beyond is to build a strong professional network that can effectively and efficiently bring together the various expertise and solutions necessary to safeguard all the digital heritage collections that are covered by UNESCO’s Charter.

10. Conclusions

We do not claim to draw general conclusions about the state of digital preservation in the Netherlands, based on the two examples discussed here.²⁷ However, the two digital archives discussed in this article are important players in the field of long-term, mass preservation of digital heritage. Their development shows that the field has made progress, beyond the phase of pilot projects that was the typical situation when the Charter was written. These examples do not allow us to conclude that we have arrived in the safe haven of ‘comprehensive and reliable preservation programmes’ that are the ideal of the Guidelines (p. 23); for this, the development of digital technology is far too rapid and volatile, the preservation community far too loose and exclusive and funding far too insufficient and unsecured. Yet, ten years of practice in developing and managing permanent large scale and complex digital archives have given the organizations involved a lot of precious knowledge and experience that could not have been achieved otherwise.

1. Archiving the products of large digitization projects has significantly contributed to the level of expertise that the KB and BenG now possess, forcing them to get a firm grip on the problem of scalability. Tackling the challenges of scalability forced them in turn to redesign their archiving systems and introduced the opportunity to achieve economies of scale and to devise new business models.
2. The high costs of mass storage and maintenance of digital content, often much underestimated by funders, are pressing the heritage institutions to devise new cost-recovery services to make ends meet. Both KB and BenG have experimented with providing paid services. The viability of

²⁷ Information on experiences gained in 20 European countries can be found in the proceedings of the Conference ‘Aligning National Approaches to Digital Preservation’ (<http://educopia.org/publications/ANADP>) held in Estonia in May 2012.

the new business models they are devising has not yet been proven, but progress is expected. The KB strives to set up an international e-Depot of scientific journals completely independent from government funding by the year 2014. BenG aims to recover some of its costs by providing archiving-as-a-service to smaller institutions.

3. We saw that both institutions struggled to emancipate themselves from an overdependence on their technology partner. What the partnership experiences with IT vendors has shown is that successful collaboration in funded research projects does not transition easily into a more business-like supplier-customer relationship. What stands out is the awareness by the institutions involved that they need to fully understand their requirements and the technology involved in order to be able to negotiate as peers. It has also become clear that the market for archiving solutions that support digital preservation is not sufficiently mature yet. The alternative choice for in-house solutions does not necessarily make the costs more controllable.
4. Conclusions (2) and (3) have important consequences for the organisation of heritage institutions and the skills required of librarians and archivists. Heritage institutions need to hire and retain highly skilled IT specialists in order to be able to carry out their tasks in a digital environment. Developing new cost-recovery services and exploiting economies of scale will require more trust-based partnerships with the suppliers of content. Trust is based on good relations and clear agreements. Cooperation between heritage institutions and building inclusive communities of practice are even more necessary in 2012 than they were ten years ago when the Charter stressed the need to 'democratise access to digital preservation techniques' (Article 11).

Acknowledgements

The writers wish to express their gratitude to Mr Hans Westerhof (Netherlands Institute for Sound and Vision), Mr Marcel Ras and Ms Astrid van Wesenbeeck (National Library of the Netherlands) for generously providing all information needed to write this article. Responsibility for all mistakes in fact or judgement rests, of course, with the authors.

Validating Quality in Large-Scale Digitization

Findings on the Distribution of Imaging Error

Paul Conway

University of Michigan

Abstract

This paper presents summary findings on the distribution of imaging error in large collections of digitized books. The paper describes a model of digitization error and presents data gathered from three statistically valid random samples of digital book-surrogates that represent large populations of source volumes digitized by Google and the Internet Archive. The findings on the frequency and severity of digitization error have important implications for usefulness of large-scale digitization. The paper concludes with observations on what wholesale digitization means for the practice of digitization. The research is funded by the US Institute for Museum and Library Services and the Andrew W. Mellon Foundation.

Author

Paul Conway is associate professor in the School of Information at the University of Michigan. His research encompasses the digitization of cultural heritage resources, the use of digitized resources by experts in a variety of humanities contexts, and the measurement of image and text quality in large-scale digitization programs. He has made major contributions over the past 30 years to the literature on archival users and use, preservation management, and digital imaging technologies. He holds a Ph.D. from the University of Michigan and is a Fellow of the Society of American Archivists.

1. The Challenge of Preserving Large Digitized Collections

From Project Gutenberg to Google Books, the large-scale digitization of books and serials is generating extraordinary collections of intellectual content whose preservation is tied to transformations in the way we read and learn. Information quality is an important component of the value proposition that preservation repositories offer their stakeholders and users.¹ For well over a decade, the cultural heritage community of libraries, archives, and museums has embraced the need for trustworthy digital repositories with the technical capacity to acquire, manage, and deliver content persistently and at scale.² Standards-based mechanisms have emerged for building,³ maintaining,⁴ and certifying⁵ preservation repositories on a scale appropriate to the preservation challenge at hand. In the new environment of large-scale digitization and third-party content aggregation, however, certification at the repository level alone may

¹ Paul Conway, "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas," *Library Quarterly* 80 (2010): 61-79.

² John Garrett and Donald J. Waters, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (Washington, DC: Commission on Preservation and Access, 1996).

³ Brian Lavoie, *The Open Archival Information System Reference Model: Introductory Guide*. Digital Preservation Coalition Technology Watch Report 04-01 (Dublin, OH: OCLC, 2004).

⁴ Priscilla Caplan and Rebecca Guenther, "Practical Preservation: The PREMIS Experience," *Library Trends* 54, no. 1 (2005): 111-124.

⁵ *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (Chicago: Center for Research Libraries and OCLC, 2007).

be insufficient to provide assurances to stakeholders and end-users on the quality of preserved content. One of the most important challenges that digital preservation repositories face is creating a capacity to validate the quality of digitized content as “fit-for use,” and in so doing provide additional investment incentives for existing and new stakeholders.

Although large-scale digitization programs have their vocal advocates,⁶ scholars, librarians, and the preservation community increasingly are raising concerns about quality and usability of image and full-text products.⁷ For example, David Bearman,⁸ Paul Duguid,⁹ and Robert Darnton¹⁰ cite scanning and post-production errors in early iterations of Google’s book digitization program. Historian Alan Gevinson surveys a collection of foundational writings on American intellectual history and finds a high proportion of minor but troublesome errors in text and illustration representation and even greater issues with the quality of underlying metadata.¹¹ Simon Tanner finds a high level of full-text error in conversion of newspapers.¹² Roger Schonfeld concludes that only the full comparison of original journal volumes with their digital surrogates is sufficient before hard copies can be withdrawn from library collections.¹³ Attempting to sort through the commentary, Daniel Cohen identifies a fundamental need for research. “Of course Google has some poor scans—as the saying goes, haste makes waste—but I’ve yet to see a *scientific* survey of the overall percentage of pages that are unreadable or missing (surely a miniscule fraction in my viewing of scores of Victorian books).”¹⁴

2. Quality in Large-Scale Digitization

The quality of digital information has been a topic of intense research and theoretical scrutiny since at least the mid-1990s. The literature on information quality, however, is relatively silent on how to measure quality attributes of very large collections of digitized books and journals, created as a combination of page images and full-text data by third party vendors. Xiaofan Lin provides an excellent review of the

⁶ Paul N. Courant, “Scholarship and Academic Libraries (and their kin) in the World of Google,” *First Monday* 11, no. 8 (August 2006), accessed August 30, 2012, http://131.193.153.231/www/issues/issue11_8/courant/index.html.

⁷ Oya Rieger, “Preservation in the Age of Large-Scale Digitization: A White Paper” (Washington, DC: Council on Library and Information Resources, 2008).

⁸ David Bearman, “Jean-Noël Jeanneney’s Critique of Google,” *D-Lib Magazine* 12, no. 12 (December 2006), accessed August 30, 2012, <http://www.dlib.org/dlib/december06/bearman/12bearman.html>.

⁹ Paul Duguid, “Inheritance and Loss? A Brief Survey of Google Books,” *First Monday* 12, no. 8 (August 2007), accessed August 30, 2012, http://firstmonday.org/issues/issue12_8/duguid/index.html.

¹⁰ Robert Darnton, “Google and the New Digital Future,” *The New York Review of Books* 56, no. 20 (2009), accessed August 30, 2012, <http://www.nybooks.com/articles/23518>.

¹¹ Charles Henry and Kathlyn Smith, “Ghostlier Demarcations: Large-scale Text Digitization Projects and their utility for Contemporary Humanities Scholarship,” in *The Idea of Order: Transforming Research Collections for 21st century Scholarship* (Washington, DC: Council on Library and Information Resources, 2010), 106–115. (See also the supplemental online report and data by Alan Gevinson (2010) “Results of an Examination of Digitizations [sic] of Books in the Field of American Intellectual History: Summary, Results, Data,” accessed August 30, 2012, <http://www.clir.org/pubs/abstract/pub147abst.html>).

¹² Simon Tanner, Trevor Munoz, and Pich Hemy Ros, “Measuring Mass Text Digitization Quality and Usefulness,” *D-Lib Magazine* 15, no. 7/8 (July/August 2009), accessed August 30, 2012, <http://www.dlib.org/dlib/july09/munoz/07munoz.html>.

¹³ Roger Schonfeld and Ross Housewright, *What to Withdraw? Print Collections Management in the Wake of Digitization* (New York, NY: Ithaka, 2009).

¹⁴ Daniel Cohen, “Is Google Good for History?” *Dan Cohen’s Digital Humanities Blog*, 12 Jan. 2010, accessed August 30, 2012, <http://www.dancohen.org/2010/01/07/is-google-good-for-history/>.

state of digital image analysis (DIA) research within the context of large-scale book digitization projects¹⁵ and establishes a “catalog of quality errors,” adapted from David Doermann and his colleagues.¹⁶ His research is most relevant because it distinguishes errors that take place during digitization [e.g., missing or duplicated pages, poor image quality, poor document source] from those that arise from post-scan data processing [e.g., image segmentation, text recognition errors, and document structure analysis errors]. Lin recognizes that, in the future, quality in large-scale collections of books and journals will depend on the development of fully automated analysis routines, even though quality assurance today depends in large measure upon manual visual inspection of digitized surrogates or the original book volumes.¹⁷

Besiki Stvilia builds on the commonality that exists in information quality models, and focuses special attention on the challenge of measuring the relationship between the attributes of information quality and information use.¹⁸ In adopting the marketing concept of “fitness for use,” he recognizes both the technical nature of information quality and the need to contextualize “fitness” in terms of specific uses. Stvilia establishes and tests a useful taxonomy for creating quality metrics and measurement techniques for “intrinsic qualities” (i.e., properties of the objects themselves). In the context of digitization products, intrinsic quality attributes are objectively determined technical properties of the digitized volume, derived from the results of digitization and post-scan image processing. By distinguishing measurable and relatively objective attributes of information objects from the usefulness of those objects, Stvilia establishes a viable research model that can be applied to the measurement of the quality of digitized books within particular use-cases.

The underlying research is part of a major ongoing research project, “Validating Quality in Large-Scale Digitization,” that is exploring the relationship between image quality (or its absence in the form of unacceptable error) and usability of digitized books. For this research project, we define quality as the absence of errors in scanning and post-scan processing relative to expected uses.¹⁹ Within the context of a large-scale preservation repository, the research adapts Stvilia’s model of intrinsic quality attributes,²⁰ and Lin’s framework of errors in book surrogates derived from digitization and post-scan processing.²¹ The overall design of the three-year research project consists of three overlapping investigative phases. Phase one defines and tests a set of error metrics (a system of measurement) for digitized books and journals. Phase two applies those metrics to produce a set of statistically valid measures regarding the patterns of error (frequency and severity) in multiple samples of volumes drawn from strata of HathiTrust. Phase three (ongoing) will engage stakeholders and users in building, refining, and validating the use-case

¹⁵ Xiaofan Lin, “Quality Assurance in High Volume Document Digitization: A Survey,” in *Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL’06)*, 27-28 April 2006, Lyon, France, pp. 319-326.

¹⁶ David Doermann, Jian Liang, and Huiping Li, “Progress in Camera-Based Document Image Analysis,” in *Proc. Seventh International Conference on Document Analysis and Recognition (ICDAR’03)*, 3, no. 6 (2003), 606-616.

¹⁷ Frank Le Bourgeois, et al. “Document Images Analysis Solutions for Digital Libraries,” in *Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL’04)*, 23-24 Jan. 2004, Palo Alto, California, pp. 2-24.

¹⁸ Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith, “A Framework for Information Quality Assessment,” *Journal of the American Society for Information Science and Technology* 58, no. 12 (2007): 1720-1733.

¹⁹ Paul Conway, “Archival Quality and Long-term Preservation: A Research Framework for Validating the Usefulness of Digital Surrogates,” *Archival Science* 11, no. 3 (2011): 293-309. [doi:10.1007/s10502-011-9155-0]

²⁰ Stvilia et al., “Information Quality Assessment.”

²¹ Lin, “Quality Assurance.”

scenarios that emerge from the research findings. More information about the design, implementation, and findings of the study is available at the project website.²²

The research project utilizes content deposited in the HathiTrust Digital Library, which is a digital preservation repository launched in October 2008 by a group of research universities, including the Committee on Institutional Cooperation [the Big Ten universities and the University of Chicago] and the University of California system. At present [August 2012] HathiTrust consists of 10.4 million digitized volumes ingested from multiple digitization sources (primarily Google). HathiTrust is supported by base funding from its 66 institutional partners, and its governing body includes top administrators from libraries and information offices at investing institutions.²³ HathiTrust is a large-scale exemplar of a preservation repository containing digitized content; 1) with intellectual property rights owned by a variety of external entities; 2) created by multiple digitization vendors for access; and 3) deposited and held/preserved collaboratively. The findings of the research are broadly applicable to the challenges in duplication, collection development, and digital preservation that are common to all digital libraries.

3. Error Model

The keystone of the study is a three-tiered hierarchical error typology and associated data variable definitions. The error model in **Table 1** identifies error at the data, page, and volume levels and establishes hypotheses regarding the cause of each error (book source, scanning, post-scan manipulation). Data and page-image errors are individually identifiable errors that affect the visual appearance of single bitmap pages. A particular error may be confined to a single page or repeated across a sequence in a volume. Whole volume-level errors apply to structural issues surrounding the completeness or accuracy of the volume as a whole, such as missing pages, duplicate pages, and ordering of pages. The development process for the error model was deeply iterative and involved substantial testing of individual error items and the meaning of narrative error definitions. The goal was to create a validated error model with clearly defined errors that could be repeatedly and consistently identified by coding staff in multiple settings.

The error model implies causality between the manifestation of observable error and one of three factors: 1) the physical qualities of the source volume; 2) the cluster of scanning activities that create a master bitmap image of two pages in an open book; or 3) the suite of post-scan manipulation processes that produce the final deliverable image that users consult. One of the primary objectives of the data collection process is to gather data on errors without assuming the cause of error. Coders were trained to “code what you see” rather than speculate on the cause of error. Data analysis then establishes cause, a topic that is beyond the scope of this paper, which focuses on the distribution of error in random samples from very large populations of digitized volumes.

²² Validating Quality in Large-Scale Digitization: accessed August 30, 2012, <http://hathitrust-quality.projects.si.umich.edu/>.

²³ Jeremy J. York, “This Library Never Forgets: Preservation, Cooperation, and the Making of HathiTrust Digital Library,” in *Proc. IS&T Archiving 2009, Arlington, VA*, pp. 5-10. See also: Jeremy J. York, “Building a Future by Preserving Our Past: The Preservation Infrastructure of HathiTrust Digital Library,” *76th IFLA General Congress and Assembly, 10-15 August 2010, Gothenburg, Sweden*.

Table 1. Model of error in large-scale digitization.

<i>Level of Abstraction</i>	<i>Possible Cause of Error</i>
LEVEL 1: DATA/INFORMATION	
1.1 Text: thick text [fill, excessive]	Source or post-processing
1.2 Text: broken text [character breakup]	Source or post-processing
1.3 Illustration: scanner effects [moiré patterns, gridding]	Scanning or post-processing
1.4 Illustration: tone, brightness, contrast	Scanning, post-processing, or source
1.5 Illustration: colour imbalance, gradient shifts	Scanning, post-processing, or source
LEVEL 2: ENTIRE PAGE	
2.1 Blur [distortion]	Scanning or source
2.2 Warp [text alignment]	Post-processing
2.3 Skew [page alignment]	Scanning, post-processing, or source
2.4 Crop [gutter, text block]	Source or post-processing
2.5 Obscured [portions not visible]	Scanning or post-processing
2.6 Colorization [text bleed, low contrast]	Source or post-processing
LEVEL 3: WHOLE VOLUME	
3.1 Fully obscured [foldouts]	Scanning
3.2 Missing pages [one or more]	Original source or scanning
3.3 Duplicate pages [one or more]	Original source or scanning
3.4 Order of pages	Original source or scanning
3.5 False pages [not part of original content]	Scanning or post-processing

The research team developed a severity scale for each of the eleven page-image errors to capture a more granular rating of each error. To train coding staff to assign severity uniformly and consistently, the research team outlined four main definitions for coders to reflect upon when assigning severity: original content, error, reading ability, and inference. *Original Content* is defined as the text or image content on the page created through the original printing process. Original content excludes marginalia, annotations, and other library-added content (bar codes, call numbers, book plates, circulation aids) added by users after the acquisition of the volume by the library. *Error* is defined as variations from the expected appearance of Original Content. *Reading ability* is designated as the ability of a reviewer to interpret the letters, illustrations, and other information contained in the Original Content of a page. *Inference* is the degree to which an average reviewer cannot detect Original Content, but must use contextual information to determine letters, words, or other information that compose the Original Content. Using this understanding, the coder is expected to apply a level of severity from zero to five for all errors detected on the page upon review. **Table 2** displays the operative severity scale used by the 12 part-time coders working in teams at the University of Michigan and the University of Minnesota.

4. Measuring Digitization Error

The research hypothesizes a state of image and text quality in which digitized book and serial benchmark-volumes from a given vendor are sufficiently free of error such that these benchmark-surrogates

Table 2. Severity scale for rating page-level error.

0 - Default - Error is undetectable on the page. ...
...
1 - Error exists but has a negligible effect on the Original Content. Show more...
2 - Error clearly alters appearance of Original Content, but has a negligible effect on reading ability. Show more...
3 - Error clearly alters appearance of Original Content and has a clear negative impact on reading ability. Show more...
4 - Nearly unable to decipher Original Content in affected area of the page; significant inference required by reviewer to obtain legibility and meaning. Show more...
5 - Original Content in affected area of the page cannot be unambiguously deciphered.

can be used nearly universally within the context of specific use-case scenarios. In the development phase, the research explored how to specify the gap between benchmark and digitized volumes in terms of detectable error. The project has developed a reliable and statistically sound data gathering and analysis system to measure error-incidence in HathiTrust volumes.

The project utilized a two-tier sampling methodology to select a sample of digitized volumes for analysis of a sample of page-images within each volume. Under the direction of the team statistician, the programmer developed a systematic random sampling algorithm to identify 1,000 volumes from the HathiTrust Library according to design-driven sampling parameters. Within each 1,000 volume sample, the project team extracted a systematic random sample of 100 pages within each volume to predict the distribution of error within the volume as a whole. The sampling algorithm is applied to the image sequence number, the complete set of which serves as a proxy for the total number of pages in a given volume, cover to cover. The algorithm divides the total number of images within a volume by one hundred to establish a number that determines the sequential sampling interval value. A random number generator establishes where in the volume (between sequence number 1 and 10) to begin sequential sampling. This method ensures that the sample will be representative of the images at the front and ends of the volumes. Sequential sampling then selects pages according to the sampling interval value, rounded up or down accordingly, to determine which whole-sequence-number image should be chosen. **Table 3** summarizes the sampling criteria, sample size, estimated population size, and number of page reviewed in the four samples drawn for the study. The number of pages-images reviewed in each sample varied slightly due to the presence to varying degrees of books with fewer than 100 pages. In general, coders reviewed approximately 25 percent of the pages in a given volume in the sample.

The research team established two data gathering groups: four part time staff at the University of Minnesota; and between four and eight part time staff at the University of Michigan. The project manager developed training materials and a training routine to establish a consistent pattern of review behavior. Data gathering commenced with sampled page-images within a digitized volume, followed by physical review of sampled volumes, and culminated in a whole volume review of the same sampled volumes from the first two samples. The scope of the project included review of 356,217 individually sampled

pages from four distinctive samples, plus a second-stage review of entire volumes totaling 691,972 page-images. Data gathering for the project took 10 months to complete. Data analysis is ongoing.

Table 3. Sample selection for page and volume review.

Sample Name	Criteria for Sample Selection	Population Size	Number of Volumes Reviewed	Number of Pages Reviewed	Whole Volume Pages Reviewed
Google pre-1923	Google-Digitized, Publication Date \leq 1923, English Language	1.3 Million Volumes	1,000	93,858	397,467
Google post-1922	Google-Digitized, Publication Date > 1922, English Language, Monograph	6.5 Million Volumes	1,000	86,439	294,505
Internet Archive	Internet Archive Digitized, Publication Date \leq 1923, English Language, Monograph	850,000 Volumes	1,000	84,539	
Non-Roman Scripts	Non-Roman Language/Script Digitized Content in HathiTrust; 4 Language/Script Categories: Arabic, Asian, Cyrillic, Hebrew	1.29 Million Volumes	1,000	91,381	

5. Findings on Distribution of Error

The findings presented here focus on the distribution of digitization error at the page-level and whole volume level for three 1,000 volume samples representing digitization activities by two vendors: Google Books and the Internet Archive.

5.1 Volumes Digitized by Google

The project utilized two independent random samples to assess the distribution of digitization error in Google Books. The breakpoint for the two samples is books published prior to and since 1923, which is the generally acknowledged dividing line in copyright for published works. This date is somewhat arbitrary, as volumes published after 1923 may be in the public domain (e.g., government publications, or author-released content), and occasionally a volume published before 1923 may remain in copyright.

Table 4 presents the absolute percentage of error across the five levels of the severity scale for the five most common page-level errors present in Google digitized books. For any given severity level, the left column is the percentage of observed error in books published before 1923, whereas the right column is the same information for books published from 1923. A severity level of zero (0) indicates that coders detected no observable error.

Table 4. Comparison of most frequent errors in Google digitized books.

		Observed Severity of Error											
		Severity = 0		Severity = 1		Severity = 2		Severity = 3		Severity = 4		Severity = 5	
		<<<< 1923 >>>>		<<<< 1923 >>>>		<<<< 1923 >>>>		<<<< 1923 >>>>		<<<< 1923 >>>>		<<<< 1923 >>>>	
Error Type	Text												
	Thick	62.04%	67.52%	25.66%	21.00%	10.73%	8.50%	1.26%	2.15%	0.19%	0.40%	0.11%	0.42%
	Broken	61.00%	73.37%	29.96%	19.18%	7.59%	5.41%	1.00%	1.27%	0.19%	0.41%	0.25%	0.36%
	Page												
	Crop	99.37%	98.85%	0.27%	0.70%	0.15%	0.13%	0.07%	0.03%	0.02%	0.04%	0.15%	0.25%
	Warp	29.22%	45.78%	60.18%	48.93%	10.17%	4.75%	0.35%	0.43%	0.04%	0.04%	0.05%	0.06%
	Obscure	16.88%	56.83%	78.05%	41.69%	4.13%	1.24%	0.40%	0.05%	0.08%	0.02%	0.46%	0.16%
Observed errors (pre-1923)		182,205		30,747		2,884		490		972			
Percent of total error		96.9%		91.2%		85.9%		82.5%		87.9%			
Observed Errors (post-1923)		113,682		17,306		3,407		795		1,077			
Percent of total error		90.5%		84.2%		85.4%		87.9%		86.5%			

For books digitized by Google, the five most common errors are the following:

- Thickened text that may range from minor aesthetic inconsistencies with no impact on readability (severity level =1) to text that is unreadable and indecipherable due to what appears to be excessive bolding. **Figure 1** is an example of thickened text at severity level 4.

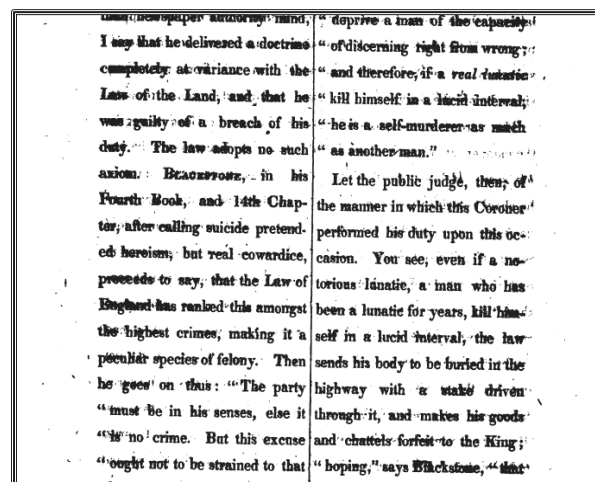


Figure 1. Thick text in Google digitized volume.

- Broken text that appears thin or sometimes nearly disappears from the page image, rendering the content unintelligible. Figure 2 is an example of broken text at severity level 5.

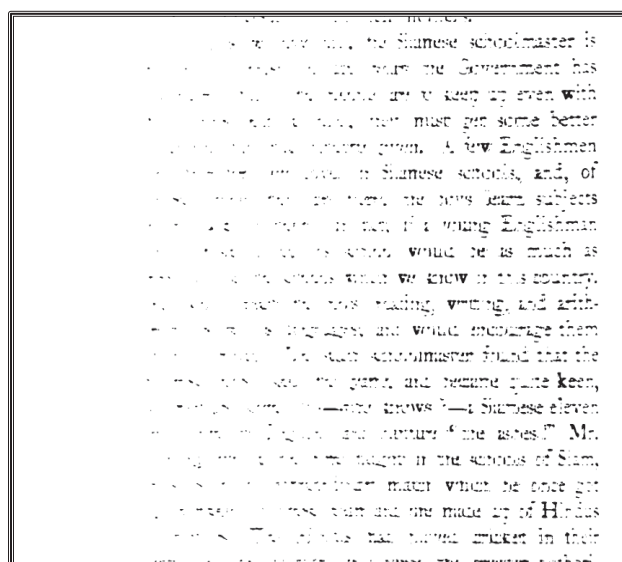


Figure 2. Broken text in Google digitized volume.

- Cropped text along the outside of the text block or at the gutter where the book is bound. Rarely, cropping may take place above or below the text block. **Figure 3** is an example of cropped text at severity level 5.

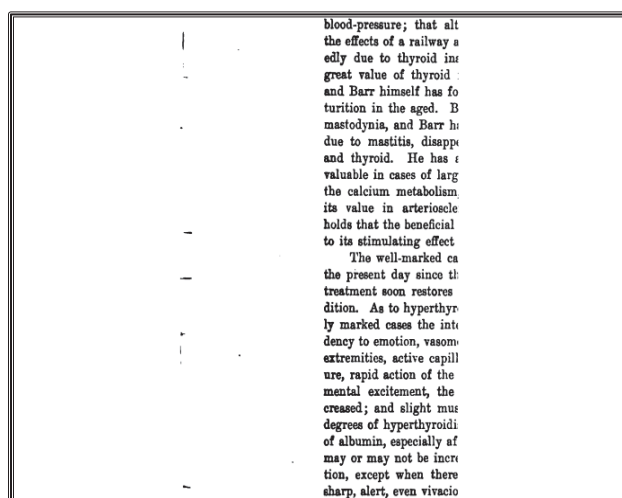


Figure 3. Cropped page in Google digitized volume.

- Warped pages feature visually distorted text blocks that may result from page movement during digitization or subtle variations in page-flattening algorithms applied after scanning. **Figure 4** is an example of a severely warped page-image.

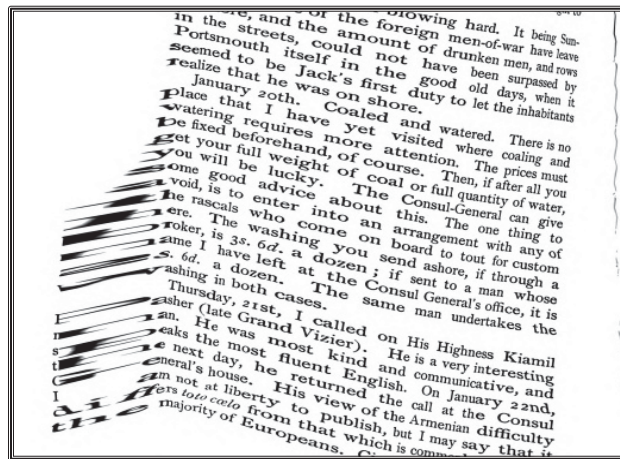


Figure 4. Warped pages in Google digitized volume.

- Obscured content due to one of two independent causes. The first reason for obscured content is the failure of the scan operator to open pages with folded content, such as large maps or graphics. **Figure 5** is an excerpt from a page with a fold-out that has not been scanned fully. A second reason for obscured content may be obstruction during the scanning process (fingers and clamps), post-scan cleanup routines, objects left in the book during scanning, such as notes or bookmarks. **Figure 6** is an example of obstruction that affects the content of the digitized page.

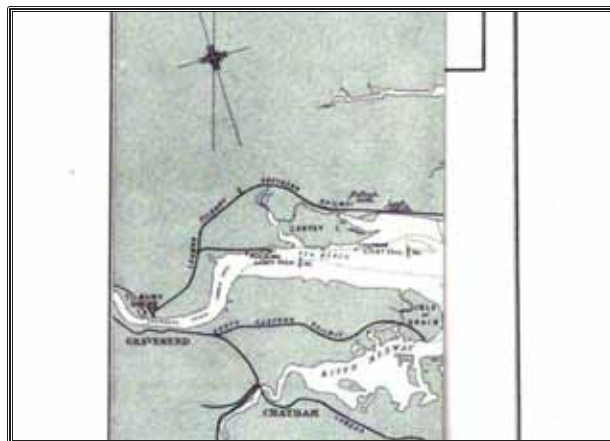


Figure 5. Obscured content (foldout) in Google digitized volume.

The distribution of error in Google digitized books yields a number of preliminary observations. First, the five most common digitization errors account for the overwhelming majority of observed error at the page image level. For example, the five common errors represent 96.6 percent of all observed error at severity level 1 in books published prior to 1923. The five errors account for no less than 82.5 percent of any given severity level (pre 1923, severity level 4). Second, for each of the five common errors, the severity of observed error declines precipitously over the range of the scale. While half of the page-images coded (48.9%) exhibit warp at severity level 1, extremely severe warp is exceedingly rare (0.05%). Third, the proportion of errors that severely compromise readability and usability (severity levels 4 and 5) is very

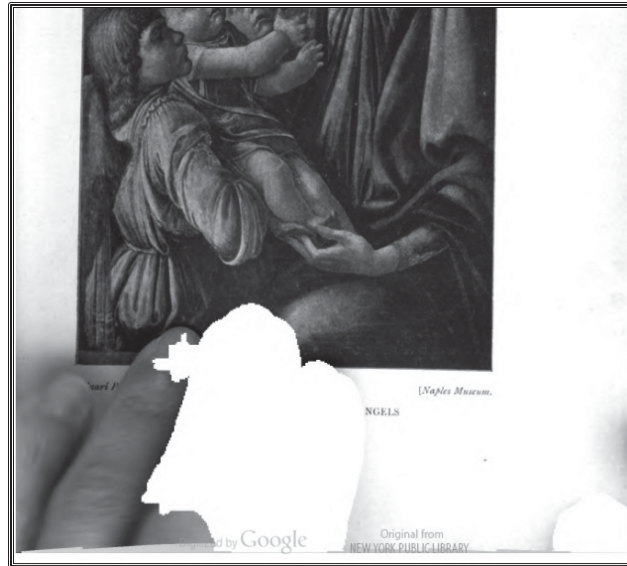


Figure 6. Obscured content (fingers and post-scan manipulation) in Google digitized volume.

small but the absolute numbers of such pages in the HathiTrust corpus is still quite large. For example, just one-quarter of one percent of page images are observed with text block cropping at severity level 5 for books published after 1922. Yet, given the scale of the population sampled (6.5 million volumes representing 1.9 billion page images) the findings suggest that almost 500,000 severely cropped images page images may be preserved in HathiTrust.

5.2 Coincidence of Error

It is possible for trained coders to observe more than one error on a given page-image. For example, the block of text could be severely cropped simultaneous with the page-image appearing to be subtly warped. The data gathering system allows coders to indicate the presence and severity of multiple errors on a single page-image. An analysis of co-occurrence of pairs of page-level errors yielded no statistically significant findings, with the exception of one text representation issue. For Google digitized volumes, broken text and thick text tend to be observed on the same page. These two errors also happen to occur most frequently at low severity levels. Sometimes variation between thick and broken text runs from the top to the bottom of the page in horizontal bands; sometimes the two errors appear randomly across the page, while occasionally, variation of text emphasis is expressed diagonally across the page. This report is not yet prepared to offer an explanation for such variation, but it is clear to the researchers that the source of such textual representation issues is not to be found in the original books. Text breakup and thickening is likely to be an artefact of a variety of post-scan transformation processes.

5.3 Consistency of Error within Volumes

Sequential sampling of page-images in a given digitized volume allows for an examination of the variation that may occur in the quality of scanning from the front to the back of a physical book. It is reasonable to expect, for example that large books, heavily paginated volumes, or books tightly bound may be more

difficult to handle and scan than books without these characteristics. For example, the appearance of warped page images might be more pronounced in the middle of a volume than at either ends.

A preliminary analysis of page images aggregated by volume and processed sequentially yielded subtle variation in error incidence across the length of digitized volumes. Such cross-volume variation is not statistically significant. For volumes published prior to 1923, broken and thickened text is observed slightly less frequently in the first fifth of sampled pages than in the remaining four-fifths of the volume. Similarly, the appearance of warped page-images is slightly less frequently observed in the front and back fifths of the digitized volumes than in the middle of the volume. These findings are preliminary in nature and must be parsed in terms of the physical and bibliographic characteristics of the digitized source volumes.

5.4 Volumes Digitized by Internet Archive

The project utilized a single 1,000 volume sample of volumes digitized by the Internet Archive (IA) and deposited in the HathiTrust Digital Library. As a matter of policy, Internet Archive digitizes only books that are in the public domain. The IA collection in HathiTrust consists overwhelmingly of volumes published prior to 1923, although occasionally volumes published more recently are included, if their contents are in the public domain.

Table 5 presents truncated data on observed error in Internet Archive digitized volumes, eliminating severity levels 2 and 3 for ease of interpretation. The table presents the absolute number of

Table 5. Digitization errors in Internet Archive digitized books.

		Perceived Severity of Error							
		Sev = 0		Sev = 1		Sev = 4		Sev = 5	
		Total	Percent	Total	Percent	Total	Percent	Total	Percent
Error Type	Text								
	<i>Thick</i>	78,819	93.23%	3,483	4.12%	6	0.01%	2	0.00%
	<i>Broken</i>	69,076	81.71%	9,952	11.77%	158	0.19%	81	0.10%
	Illustration								
	<i>Tone</i>	58,329	69.00%	20,835	24.65%	146	0.17%	10	0.01%
	Page								
	<i>Blur</i>	79,654	94.22%	3,618	4.28%	34	0.04%	27	0.03%
	<i>Warp</i>	34,772	41.13%	48,184	57.00%	2	0.00%	0	0.00%
	<i>Crop</i>	84,211	99.61%	187	0.22%	16	0.02%	63	0.07%
	<i>Obscure</i>	48,127	56.93%	33,487	39.61%	17	0.02%	60	0.07%
	<i>Colorization</i>	40,204	47.56%	38,230	45.22%	4	0.00%	23	0.03%
	<i>Skew</i>	76,291	90.24%	7,890	9.33%	0	0.00%	0	0.00%
Total Error		569,483		165,866		383		266	
Observed Errors (Google)				95,293		199		206	
Proportion of total error				57.5%		52.0%		77.4%	
Observed errors (IA)				70,573		184		60	
Proportion of total error				42.5%		48.0%		22.6%	

observed error and proportion of error represented by severity levels 1, 4, and 5. Severity level zero (0) is the default value of “no observable error.” Error at severity levels 4 and 5 compromises the readability and usability of the page image, whereas severity level 1 represents detectable error at the lowest level, such that readability is not affected. The table is also annotated to indicate four types of error that occur relatively frequently in Internet Archive digitized volumes but that are rarely observed in Google digitized volumes.

For Internet Archive volumes, these four errors are:

- Tonal qualities of the overall page image that effectively reduce the contrast between text and background. Tone issues lend a general muddy complexion to the page image and may range in severity from minor to complete loss of readability. **Figure 7** is an example of tonal error at severity level 2.

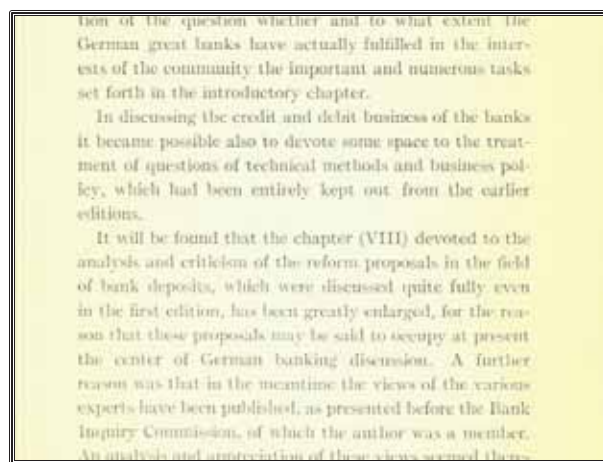


Figure 7. Tonal error in Internet Archive.

- Colorization errors in applying an artificial ageing mask to the page image. The net result of this error is extremely false colour values for a page image, ranging from phosphorescent yellow through pumpkin orange. Accompanying colorization errors, applied as a post-scan processing step, may be broken text, blur, and other textual artefacts. **Figure 8** is an example of colorization error at severity level 5.

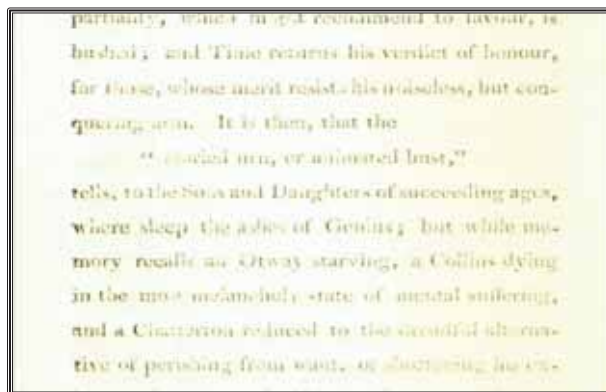


Figure 8. Colorization error in Internet Archive.

- Skewed page images occur when the source volume is placed under a digital camera at an angle that cannot be automatically corrected with deskewing software or when such software has not been used effectively in post-scan processing. **Figure 9** is an example of a skewed page image at severity level 1, assigned to the image because no content is lost in the process.

Phelps v. Pitcher	409	Poplin v. Hawke	539
Phelps v. Schultorpe	297	Porter v. Byrne	275
Platt v. McCullough	29	Porter v. Ferguson	308
Pickard v. Bailey	486, 514	Porter v. Johnson	113
Pickard v. Sears	294	Porter v. Pillsbury	322
Pickering v. Bp. of Ely	115	Porter v. Poquonnoe Man. Co.	440
Pickering v. Dowson	281	Porter v. Seiler	54
Pickering v. Noyes	246	Porter v. State	432
Piton's (Gen.) case	492	Porter v. Baker	532
Piddock v. Brown	361	Porter v. Ware	386
Pierce v. Butler	399, 401	Porter v. Webb	55
Pierce v. Chase	423	Potts v. Everhart	109
Pierce v. Hoffman	53	Poulney v. Ross	118
Pierce v. Parker	288	Poulter v. Killingbeck	271
Pierce v. Weymouth	304	Powell v. Hord	394
Pierce v. Wood	112	Powell v. Millsburn	35
Pierson v. Hutchinson	558	Powell v. Monson	26
Pigot v. Davies	521	Powell v. Blackett	572
Pigot v. Holloway	437	Powell v. Bradbury	473, 559
Pike v. Crechore	513	Powell v. Edmunds	281
Pike v. Hayes	109	Powell v. Ford	577
Pile v. Benham	428	Powell v. Gordon	392
Pim v. Currell	139	Powell v. State	462
Pipe v. Steel	356	Powell v. Waters	164
Pitman v. Maddox	117	Power v. Frick	576
Pitt v. Chapelow	297	Power v. Kent	239
Pitt v. Shaw	49	Powers v. McFerran	575
Pittam v. Foster	176	Powers v. Nash	188
Pittam v. Walter	510	Powers v. Russell	74
Pittsfield, &c. P. R. Co. v. Harri-	510	Powers v. Shepard	323
son	484	Powers v. Ware	566
	31, 37	Powers v. Johnson	116, 120

Figure 9. Skew error in Internet Archive.

- Blur is an artefact of camera or object motion during the scanning process or the result of lighting anomalies from the plate glass covering the book during scanning. **Figure 10** is an example of the blur error at severity level 1.

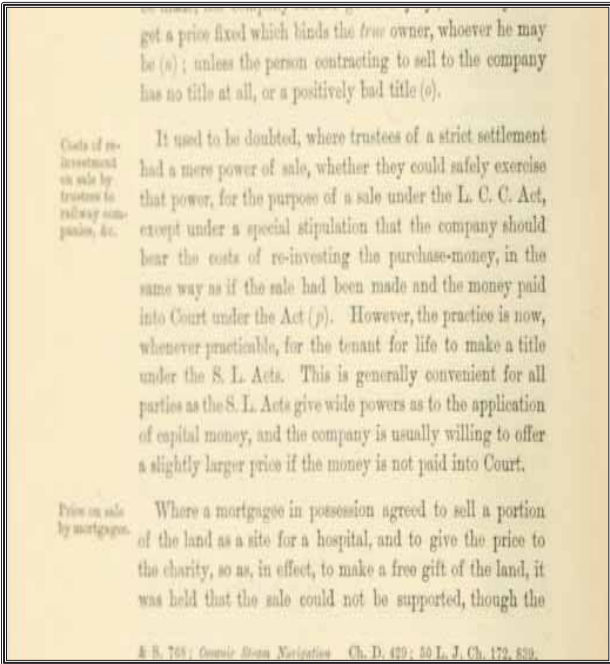


Figure 10. Blur error in Internet Archive.

Table 5 demonstrates the significant differences in patterns of error between volumes digitized by Internet Archive and Google. Several preliminary observations are in order. First, the absolute number and proportion of severe error is very low to non-existent in Internet Archiving. Only 649 errors at severity levels 4 or 5 (0.7%) were observed through the careful inspection of 85,535 page-images. This proportion is well within the statistical limits of accuracy for the error observation process. Second, at the low end of the severity scale, different errors dominate Internet Archive digitized books. The five errors that account for over 90% of the error observed in Google digitized books account for only 57% of the errors observed in Internet Archive collections. For IA, scanning and post-scan processing create distinctive artefacts, including the appearance of aging, blurring of text, low contrast text representation, and a sufficient level of skewing to be noticeable to the casual observer. It is outside the scope of this paper to speculate on the causes of these not-so-subtle differences in the look and feel of Internet Archive digitization.

5.5 Severe Error in Volumes

Tables 4 and 5 reported the distribution of digitization error within the page-image set of independent samples of digitized volumes without regard to the patterns of error within discrete volumes. In reality, digitization error is manifested not within a collection of page images but rather as problems associated with an individual book-surrogate. Digitization quality is not necessarily a randomized occurrence across a set of page-images but rather is a problem that may affect some digitized volumes more severely than others. The analysis of digitization error thus must be evaluated at the level of aggregated volume, where a random sample of page images associated with a given volume are analysed together and compared across the total number of volumes within a given sample.

Table 6 is a first effort to present a portion of this analysis of volume-specific error. For each of the three samples under consideration in this paper, the table presents the total number of volumes in the sample with pages coded with errors of any type coded at severity levels 4 or 5. At these severity levels, readability is compromised to the point that the volume may be unusable. For Google digitized volumes, books published prior to 1923 (<<<<) are in the left column, while post-1922 books (>>>>) are in the right column. The table then extends this data to show the proportion of volumes with compromised pages and the cumulative percent as the number of compromised pages increases from a minimum of 1 per volume to the maximum observed compromised pages. The total number of volumes analysed in any given sample is less than the 1,000 drawn volumes because the analysis excluded volumes with less than 50 page-images.

A number of striking observations are possible from even a casual reading of the table. First, for Google digitized volumes, between 59% and 69% of all digital surrogates have no observable severe error. Volumes published after 1922 are less prone to severe error than older volumes. Second, the proportion of digital volumes free of severe error is significantly greater for Internet Archive digitized volumes. In the study, fully 93% of all volumes inspected have no severe error. When considering volumes with only one page of error at severity levels 4 or 5, nearly 98.5 percent of Internet Archive digitized volumes are error free, whereas the proportion of Google volumes with one or fewer severe errors ranges from 77% to 83%.

At the other end of the error spectrum, Table 6 exposes the existence of what might be called “bad books”: digitized volumes with more than 10 pages per volume with severe error of any type. For the Internet Archive sample, only 4 of 940 volumes or less than ½ of one percent (0.4%) are rife with error laden pages. For the Google samples (pre-1923 published and post-1922 published), the numbers of bad

Table 6a. Severe error in Google digitized volumes.

Pages with Severe Error in a Volume (<<<< pre 1923 post- 1923 >>>>)							
Pages w/ Severe Error		Number of Volumes		Proportion of Sample		Cumulative Percent	
<<<< 1923 >>>>		<<<< 1923 >>>>		<<<< 1923 >>>>		<<<< 1923 >>>>	
0	0	555	637	59.55%	69.16%	59.55%	69.16%
1	1	167	131	17.92%	14.22%	77.47%	83.39%
2	2	76	50	8.15%	5.43%	85.62%	88.82%
3	3	39	29	4.18%	3.15%	89.81%	91.97%
4	4	24	11	2.58%	1.19%	92.38%	93.16%
5	5	12	8	1.29%	0.87%	93.67%	94.03%
6	6	12	6	1.29%	0.65%	94.96%	94.68%
7	7	8	3	0.86%	0.33%	95.82%	95.01%
8	8	6	3	0.64%	0.33%	96.46%	95.33%
9	9	1	2	0.11%	0.22%	96.57%	95.55%
10	10	4	3	0.43%	0.33%	97.00%	95.87%
11 to 21	11 to 28	20	28	2.00%	3.10%	99.10%	98.97%
22 to 68	38 to 168	8	10	1.00%	1.00%	100.00%	100.00%
		932	921				

Table 6b. Severe error in Internet Archive digitized volumes.

Pages with Severe Error in a Volume (pre -1923)			
Pages w/ Severe Error	Number of Volumes	Proportion of Sample	Cumulative Percent
0	876	93.19%	93.19%
1	43	4.57%	97.99%
2	7	0.74%	98.51%
3	2	0.21%	98.72%
5	3	0.32%	99.04%
6	3	0.32%	99.36%
8	2	0.21%	99.57%
11	1	0.11%	99.68%
26	1	0.11%	99.79%
27	2	0.21%	100.00%
	940		

books are also small but constitute a larger proportion of the whole than is the case for the Internet Archive. More recently published volumes are more severely error prone than older volumes. Thirty-eight volumes (4.1%) published since 1922 have 11 or more severe errors, while 28 volumes in the earlier sample (3.0%) are similarly afflicted. Although these numbers are small in proportion, when the sample is projected to the full population of digitized volumes, the absolute number of corrupted and possibly unusable digital surrogates may present a significant barrier to use of the collection as a whole.

5.6 Volume Level Error

To this point in this paper, the findings presented are derived from a manual page-by-page review of a sample of pages from a sample of digitized volumes. This two-tier sampling strategy is incapable of detecting and measuring errors that pertain to the volume as a whole. The error model developed for the project allows for such errors, including missing pages, apparently duplicate pages, pages in the scanned volume that appear to be out of order according to printed pagination, and false pages that do not belong with the digitized volumes. Additionally, for the Google digitized volumes, it may be possible to detect through a rapid review of the entire digital surrogate page-images that contain published foldouts that have not been scanned. The project developed an efficient review interface for measuring whole volume errors. A secondary purpose of the whole volume review process was to establish benchmarks for how rapidly manual review could take place on an entire volume with an acceptable level of review quality. It is outside the scope of this paper to report on review timing and cost findings.

Table 7 presents summary findings for the observation of whole volume errors in Google digitized books published before 1923. The 1,000 volume sample contains 397,467 page-images with possible error. In this whole volume experiment coders reviewed the entire digitized volume page by page, paying particular attention to the pagination of the volume as well as clues regarding severe error and obstructed content.

Table 7. Whole volume errors for Google digitized books (pre-1923).

pages reviewed n = 397,467							
Pages per Volume (n = 1000)				Mean	Std. Dev.	Minimum	Maximum
				397.49	272.75	8	1628
Type of Error	# Pages	Proportion					
<i>Missing Page(s)</i>	660	0.17%		0.67	6.529	0	155
<i>Duplicate Page(s)</i>	572	0.14%		0.62	4.48	0	92
<i>Pages Out of Order</i>	240	0.06%		0.24	2.185	0	43
<i>False Page (s)</i>	41	0.01%		0.04	0.343	0	8
<i>Obscured Content</i>	3307	0.83%		3.36	20.82	0	366

The results of a whole volume manual review raise a number of issues about the integrity of digitized volumes. First, the absolute number of whole volume errors is quite low relative to the overall number of page images in the 1,000 volume sample. For example, the total number of missing pages detected in the review is 660 or 0.17 percent of the total number of pages reviewed. Duplicate pages, which reflect the actions by scanning technicians to achieve a fully rendered image or compensate for glassine tissues placed in volumes by publishers to protect fragile illustrations, are also few and far between. False pages are also a rarity in volumes deposited in HathiTrust. A false page may result from a post-scan processing error or from scanning camera misalignment. Obscured content, most likely foldouts not scanned fully in the Google scanning laboratories, account for the largest number of scanning errors in the whole book review.

Against this backdrop of low incidence of error is the presence of severely disrupted surrogates, a problem made most clear in the missing pages category. Although missing pages are rare, when they do occur in a given digital surrogate, they tend to be repetitive and frequent. Given the current state of digitization processing and automated quality assurance, missing pages in a digital surrogate present one of the most problematical issues for the overall acceptance of large-scale digitization.

6. Forthcoming Analysis

Among critics, large-scale digitization has raised the specter of entire libraries of digital books whose content is unusable due to errors in imaging and metadata. At the very least, Gevinson²⁴ and the few others who have attempted systematic explorations of digitized books find the randomness of low level page-image error to be annoying—perhaps sufficiently so to reject the entire effort as technologically suspect. Other commentators apparently have based their criticisms of Google Books on impressionistic grounds, preferring to make general conclusions about the corpus of digitized books on the basis of comparison of digital surrogates of personal favorites with copies on a library’s shelf or personal collections.

This paper establishes an initial benchmark on the existence and distribution of error in digital surrogate representations of books digitized by Google and the Internet Archive. Volumes from these two third-party technology firms represent over 98% of the content in the HathiTrust Digital Library as of this writing. Further analysis of the data gathered for the study investigates the reliability of the data gathering methods and the hypotheses of causality that are embedded in the project’s error model.

Findings from page-level and volume level error will yield a prioritized list of scanning and post-scan procedures that result in error. Future research will explore the extent to which the most frequent and the most offending errors can be detected and corrected using automated image processing algorithms. Preliminary research has identified potentially valuable processing procedures for duplicate page images, and for warped or skewed page images. Fixing text anomalies might also be possible in certain cases. The challenges of correcting scanning artefacts in book illustrations are more problematical. It is unlikely that existing digital imaging analysis techniques will be able to detect all digitization errors in the HathiTrust collections, although the most common and most egregious errors could be given special priority in the development of new image processing algorithms.

A supplemental goal of the project is to address a priority need within the HathiTrust community of stakeholders: namely a tool for the efficient review of individual volumes on demand and the rating of these volumes in terms of the presence or absence of critically important errors. This work is ongoing and will become one of the principal deliverables of the grant project.

7. Implications for Practice

This research is not about digitization operations, but rather about the consequences for practice when the cultural sector does not control the terms of digitization. In a world of third party digitization and a growing network of digital preservation repositories collecting digital content that they did not create, quality becomes a property to be explored, documented, and communicated. In this context, a lowered bar for acceptable quality is not necessarily an ethical or professional compromise, but rather is a mandate

²⁴ Gevinson, “Results of an Examination.”

that must be undertaken systematically and in the context of use. Equally important is the need to communicate findings on digitization quality (or the systematic lack thereof) to end users through some form of exposed metadata.

The major contribution of the research project for the practice of digitization is to be found in the new tools and techniques for measuring quality, for analysing the data gathered systematically, and in the context that such investigations provides for judging the value of preserved digital content. The error model presented in summary in this article is specific to the digitization of books in a factory-like setting, where speed and efficiency trump any quality assurance processes that limit the productivity of the overall scanning operation. The specific model, however, may be modified and applied to the digitization of other media and information formats that are commonly found in the archives of cultural organizations, particularly typescript archival materials.

Perhaps the most important implication of this research relates to the establishment and certification of trust in digital preservation repositories. The current model for certification of trusted digital repositories is largely designed to endow trust at the organizational level and to the entirety of technical systems built to take in, manage, and deliver digital content over time. Little or no attention is devoted in the certification framework for systematic evaluation of the qualities and information value of the preserved digital content itself. The research reported here suggests a viable strategy for ongoing quality assessment. Additionally, the research suggests that even a small incidence of serious error can be problematical because of the overall message that such debilitating error sends to the end user. Finding, assessing, tagging, and communicating the existence of serious error should be a routine component of the repository certification process. Such routine quality assessment can be done manually, as has been done in this research project, but would benefit from special attention to automated image error detection routines.

Libraries, archives, and museums have been engaged in the digitization of books for over 20 years. Digitization in the cultural heritage sector has been governed for most of these decades by community-established guidelines and best practices, some aspects of which have made their way through national and international standards processes. Cultural heritage professionals are rightly wedded to the maintenance of a high bar of quality for works chosen after a carefully administered selection process for digitization in locally managed labs or highly trusted digitization vendors. In this well established, vertically integrated process, the cultural heritage sector can exercise the kind of control over workflow and end product specifications that helps guarantee image quality and a low to non-existent level of error. The price for such adherence to locally generated and managed workflow is a relatively slow pace of converting highly valuable and useful content from analogue to digital formats. The research findings reported here in no way undermine or compromise the rights and responsibilities of cultural heritage professionals to set a high bar for quality in digital reproduction.

Acknowledgements

The project involves collaboration between the University of Michigan, the University of Minnesota, and the HathiTrust Digital Library. The Andrew W. Mellon Foundation supported project planning and research design. A grant from the Institute for Museum and Library Services [LG-06-10-0144-10] supported system design, data gathering, and data analysis and reporting. The author thanks Jacqueline Bronicki, Project Coordinator; Ryan Rotter, Systems Programmer; Ken Guire, Statistician; Jeremy York, Associate Librarian; and co-PI Edward Rothman, Professor of Statistics, for their indispensable work on the project.

Lost in Transit

The Informative Capacity of Digital Reproductions

Lars Björk

National Library of Sweden

Abstract

Digitization is an act of interpretation, the relation between the source document and its reproduction is never a 1:1 situation; alterations are always introduced in the process. The characteristics of the digital reproduction results from assumptions regarding the informative potential of the source document, assumptions that guide the formation of the digitization process and decide the parameters according to which it is performed. Without knowledge of these factors it is difficult to evaluate to what extent the reproduction can fulfil its intended purpose. This situation is especially evident in relation to digitization within the library sector where the focus often is singlehandedly placed on the textual inscription of a document whereas other aspects, such as its material characteristics and sociocultural context, are given considerably less attention. The article will discuss a multidimensional approach to documents, where these aspects, rather than being isolated factors are seen as complementary components. This approach can be used to model the informative capacity of the reproduction vis-à-vis the source document, thus facilitating both the design of processes and the informed use of their products.

Author

Lars Björk, Senior Conservator at the National Library of Sweden, commenced his Ph.D study at the Swedish School of Library and Information Science, UC Borås / Gothenburg University in 2009. His is studying digitization processes within the library context, with a special focus on factors that determine the informative capacity of the reproduction in relation to the source document. Lars Björk holds a BA & MA in Paper Conservation and has been on the staff of the National Library since 1995. He was head of Conservation 1997-2005 and thereafter Preservation Coordinator. 2005-2009 he served as a standing committee member of IFLA and LIBER. He was the Chair of the Swedish section of IIC, 2004-2007.

1. Introduction

It goes without saying that digitization is an integral component of collection management in the MLA¹ world today. Documents and objects, previously accessed on location in heritage institutions, are now transferred into digital formats and engaged as binary encoded representations. The ubiquitous presence of digital technology in our everyday lives has led us to accept these representations as surrogates while the products of other reproduction processes, such as photocopies and microfilm images, always have been perceived as (more or less imperfect) supplements.

The developments within the field tend to focus on issues relating to the administration of information that is already in digital format, e.g., digital infrastructure, long-term sustainability, and the capacity to manage and process large data sets. But the relation between the source document and the reproduction—the actual transfer from the analogue to the digital domain—has garnered considerably less interest. Still,

¹ Museums, Libraries and Archives

this relation, to a large extent, determines the capacity of the reproduction to serve as a substitute for the source document or as an enrichment, a pointer, or a comment on a particular document.²

The characteristics of the reproduction always are dependent on the presuppositions regarding the informative potential of the source document. *Informative configuration* will be used in this article as a concept that defines the ways in which a document is regarded as informative. Digitization is an act of interpretation, the relation between the source document and its reproduction is never a 1:1 situation; alterations are always introduced in the process. Without knowledge of the underlying assumptions it is difficult to evaluate the informative capacity of the reproduction vis-à-vis the source document. There is an element of trust involved in relying on reproductions. The concept of reliability, as defined in the InterPARES framework, offers a useful starting point in the analysis of the factors that guide informed use of reproductions.³

The trustworthiness of a record as a statement of fact. It exists when a record can stand for the fact it is about, and is established by examining the completeness of the record's form and the amount of control exercised on the process of its creation. (InterPARES2)

I will use the concept of reliability to construe the relation between the source document and its digital reproduction.⁴ Although references will be made to the MLA-sector in general, this article will focus on digitization within the library setting. This paper, based on some of the findings from an ongoing Ph.D.-project,⁵ will start by addressing the source document and its information potential, it will then consider the document in the reproduction process. The paper concludes by suggesting a conceptual model that can be used to outline the information valence of the digital reproduction vis-à-vis the source document.

2. Scope

To what extent can we examine the completeness of the reproduction's form and the amount of control exercised on the process of its creation? The complexity and seeming opacity of digital reproduction technology often causes digitization to be regarded as a kind of black box in the Latourian meaning; an incomprehensible process where documents are entered and refined into pure information in a transformation of almost alchemical proportions. The user's ability to assess the characteristics and particulars of the reproduction process are in most cases limited as documentation regarding the processes often is missing or of a fragmentary nature, see e.g., Ross (2004) and Warwick et.al. (2009). In many instances this is not seen as a question of vital importance. Abram & Luther (2004) have used the term

² The term "*transmission*" is in this article used in reference to the transfer of information from a source document to a target document. "*Source document*" is used to avoid the ambiguous connotations that accompany the term "original" in relation to text carrying documents. "*Reproduction process*" is here understood as a subset of transmission procedures, distinguished from other methods by its capacity to reproduce a source document that exists independently of the actual process. "*Reproduction*" will be used in reference to specific products of such processes, e.g., "a digital reproduction." "*Digitisation*" refers to reproduction methods relying on digital technology to capture information content in analogue format and transform it to digital format.

³ Reliability, authenticity, and accuracy are the three factors that constitute the concept of Trustworthiness. (InterPARES2)

⁴ Authenticity and accuracy are of course also highly relevant concepts in relation to digitisation processes but they will not be discussed here.

⁵ The project studies the capacity of digitisation processes to capture and re-present printed text carrying documents (e.g., printed books) in digital format. A focus is placed on the characteristics of the reproduction and its potential to replace the need to access the source document.

“format agnostic” in reference to the growing group of users who regard the web as a primary source of information. For these users there are no differences in credibility between print media—accessed in situ at institutions such as libraries and archives—and text based resources accessed via the web; the buzzword being “*information is information.*”⁶ It has even been argued that the preference for physical source documents in the humanities might endure for completely different reasons—the scholarly community still refers to the *printed book* as the standardised proof of academic achievement, e.g., Di Leo (2010). On a similar note Ross (2004) have suggested the concept “*Presumption of Authenticity*”⁷ to describe the readiness of users of web based resources to uncritically accept that reproductions indeed represent what the producer tells they do. Paul Conway (2010) summarises this trend:

Google is a metaphor for the instant gratification expected in information search and retrieval today. For a new generation of users, Google represents anonymous access to information without human mediation. (Conway 2010, 63)

The question of trustworthiness of digital reproductions has thus gained increasing attention especially among users with an interest in assessing primary sources in heritage collections. The underlying issue is the difficulty in establishing an unambiguous relation between the source document and the reproduction. Deegan & Sutherland observes that: “...*the relationship between any original object and a reproduction is massively compromised by the sheer impossibility of exact iteration.*”⁸

In order to understand how the reproduction process effects the characteristics of the digital surrogate we need to address both the informative potential of the documents that are reproduced and the particulars of the process used to capture and present them.

3. The Shapes of Documents

Though we often regard documents as self evident, their inherent nature and their relation to the concept of information have been the subject of much debate. The theoretical approaches employed within the field of Library and Information Science (LIS) span the two positions that Furner (2004) refers to as the objectivist and the subjectivist stance, i.e., information considered either as a property *inherent* in a medium or information considered as a property *ascribed to* this medium.

Both these positions can lead to reductionist conclusions regarding the role of documents in the transfer of information. At times, the document is regarded as a passive container, a device whose informative content is seen as a discrete entity that without alteration can be moved to a new carrier of choice, e.g., the conduit model.⁹ At others, its essence is seen as emanating from material structures that bear no significance apart from that which is attributed to it by the user; see e.g., Talja et.al. (2005) for a discussion of some of these approaches. But acknowledging the role that a document can have as a necessary prerequisite for the management of information does not lead to a consensus on which factors

⁶ Abram & Luther (2004)

⁷ Ross (2004, 8)

⁸ Deegan & Sutherland (2000, 157)

⁹ The conduit model is an interpretation of Shannon’s signal theory (Shannon 1948) that describes a statistical approach to the removal of disturbances - “noise” - during transfer of signals. It has been applied metaphorically on human communication, referring to messages as entities with a fixed informative content that (under certain conditions) can flow without alteration between sender and receiver.

actually determine the nature of this document.¹⁰ The positions on this issue range from descriptions of physical properties such as a “... *more or less flat surface or on surface admitting of being spread flat when required...*” (Ranganathan 1963) to the concept of *immutable mobiles* (Latour 1986), (Brown & Duguid 1996). We also find definitions based on *institutional practices* (Hjørland 2000), *the assignment of human agency*, i.e., speech, (Levy 2003), *bibliographical practice* (Smiraglia 2001).

4. Informative Configurations

In order to avoid any conceptual drawbacks that might arise from adhering to a document definition based on specific characteristics or modes of practice, this study has been based on a more inclusive approach, one used by the French cross-disciplinary research group Roger T Pédauque. It has engaged the multifariousness of documents by outlining a set of constitutive dimensions that in combination with one another constitute the characteristics of a document (Pédauque, 2003). Similar multidimensional approaches are found both within the field of LIS, e.g., Lund (2010) and in other disciplines that address the informative potential of material artefacts; e.g., museology (Maroevic 1998), Media studies (Meyrowitz 1998) Archival theory (InterPARES2), Conservation (Caple 2000).

Pédauque (2003) defines the document as consisting of three informative dimensions:

“*FORM*” - material characteristics, the physical structure and material composition.
 “*SIGN*” - text based information content, stored in sequences of alphanumeric signs)
 “*MEDIUM*” as a means of communication and providers of information, the functions and relations that are ascribed to the document as part of a culturally established systems
 The informative dimensions of a document, after Pédauque (2003)

These dimensions are not delimited by clear, easily recognisable boundaries; they should rather be regarded as corresponding and complementary aspects that in different configurations can be used to describe how the informative capacities of documents vary with context.

Thus the notion of *informative configurations* can be used to describe the capacity of a document. For example, we can consider a commonplace document such a telephone directory. If it is used to find a specific telephone number, it will be configured by the user to focus on its function as a carrier of text based information—the dimension “*SIGN*”—(e.g., names, addresses, and telephone numbers). But the physical organisation of these signs—the dimension “*FORM*” (e.g., indexing based on alphabetical order)—focuses on the formal dimension of this information. Finally, the directory has to be socially sanctioned—the dimension “*MEDIUM*” (by being published by an authorised body; e.g., the Telecom Administration)—to be considered a reliable source of information.

If the directory instead is viewed within a socio-cultural context, e.g., in a study of societal implications of telecommunication development, the dimension “*MEDIUM*” will receive primary

¹⁰ See Francke (2005) for a general discussion of the document as a concept. The parameters of the document in digital format are as evasive and difficult to establish as in the analogue domain. Xie summarises: Thus, the concept of *document* in the digital world now has two clearly distinguishable yet interrelated dimensions: a cognitive one that still treats a document as a coherent whole and an operational one that recognizes the composition of a document’s preservable parts as required by reproduction. Both dimensions are necessary because the cognitive dimension maintains the human users’ understanding of the content and context of digital documents, and the operational dimension provides guidance for the actual preservation activities. (Xie 2011) (emphasis in original)

attention, as a: “... a vector of message between people,”¹¹ while the text based content and material characteristics will be considered as complementary aspects. We can use these dimensions in our study of the digitization of documents.

5. The Digitization Process

The digitization process is often spoken of in terms of a chain. The merit of the concept of a “*digitization chain*”¹² is that it underscores its processional character and clearly demonstrates that it consists of a series of functions, i.e., the actual information capture is only one of several connected activities. However, a weakness of the chain metaphor is its linearity.

An analysis of documentation¹³ regarding digitization reveals the process as consisting of a large number of different activities. Unlike a chain, the digitization process can be regarded as five groups of interlinked activities that can be termed: selection, information capture, processing, archiving, and presentation. These activities can be seen as representing the core functions of any digitization process. But the process consists of two, partly parallel flows of documents, i.e., that of the source documents and that of the reproductions. Furthermore, there are other types of information transfer taking place at each functional phase; e.g., production and processing of technical, administrative and structural metadata. There are also various types of files which are derived from the master file as well as production of information for the purpose of quality control and statistics.

The complexity of this process is better envisaged if it is conceptualised as consisting of five clusters of activities rather than a single linear flow (Figure. 1).

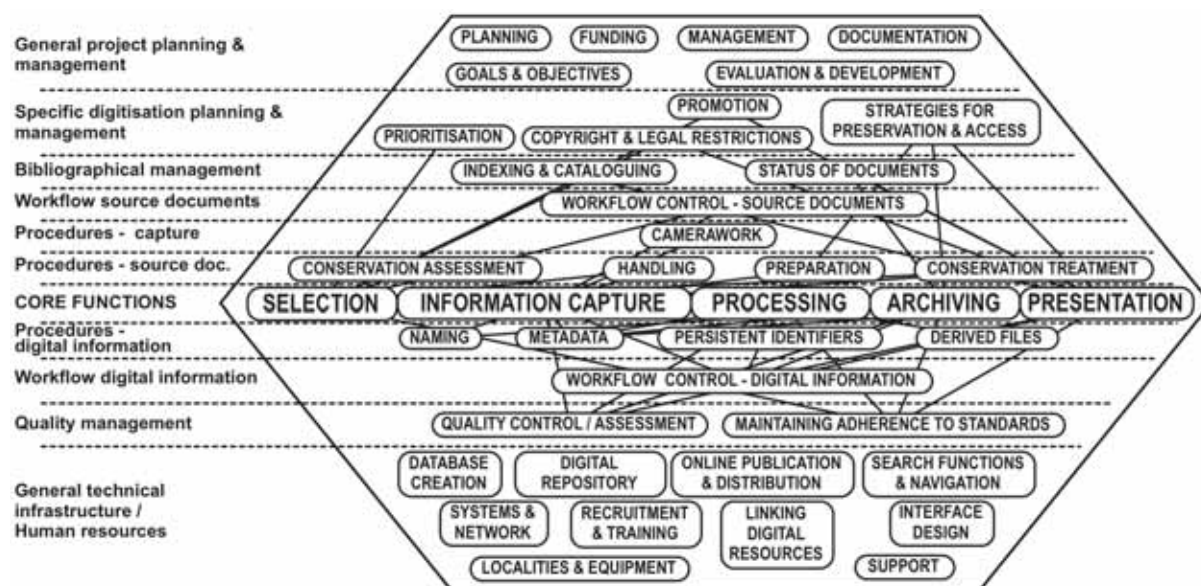


Figure 1. The “big picture” of the digitization process.

¹¹ Pédaque (2003, 17)

¹² See IFLA (2002) and Cornell Univ.

¹³ The project is partly based on an analysis of documentation and reports made by digitising institutions and organisations.

The digital surrogate should therefore more rightly be regarded as one of several informative products (albeit a primary) that are generated in the digitization process. If anything, the digitization process bears a resemblance with that which Hillesund (2006) describes as “*the text cycle*”, thus contradicting the notion of a series of functions with a clearly defined start end point. Most activities within the process are standardised and follow generally established guidelines. But there are several factors that are not as readily defined; e.g., the methodology used for selection, the procedures during capture, the extent of copy specific information processed, the bibliographical status of the reproduction and the functionality and manoeuvrability of interfaces used for searching and dissemination. These “soft factors” are of greater consequence for the status of the digitized library document than for the archival record. Digitization processes within the archival domain tend to deal with collections of records with more or less standardised formats whereas printed books generally display a larger variance with regard to physical characteristics and organisation of text based information content. Another difference is that the archival record is represented as a unique instance with regard to administrative status, provenance and content. The printed book on the other hand, being the result of a processes of mass production, generally are referred to on a higher level of abstraction. Although each digital reproduction of a library document can be traced back to a specific copy, the associated information does not necessary refer to this specific instantiation.

As the model of the digitization process indicates, the relation between the source document and the reproduction is complex since each activity in the process will have an impact on the characteristics of the final product.

6. Digital Restructuring

A distinguishing feature of digital technology as a method of reproduction is the radical reconfiguration to which it subjects the source document. The structural principles that apply to a printed document—a physical substrate organising text-based inscriptions—are recontextualised once the document is transferred into digital format. This can be described as a shift from a document where storage and display is performed in the same unit, e.g., on the pages of a printed book, to an instantiation where these two steps take place in separate units; e.g., a text in digital format where storage is assigned to a hard disc drive and the computer screen serves as the display (not necessarily in physical proximity to the storage device¹⁴).

But this split between storage and display is accompanied by a simultaneous amalgamation of the functions of medium and tool. While they represent discrete activities within the analogue context, it is not meaningful to separate them in the digital domain. The computer is at the same time both a writing tool and a storage device.¹⁵ Another characteristic aspect of the digital transformation concerns the modes in which text based information is coded during storage and display.¹⁶ In the case of a printed book, the letters on the pages function both as storage and presentation signs, whereas digitally stored text has to be decoded—from storage signs (binary code stored on e.g., a hard drive or a flash device) to presentation sign (e.g., alphanumeric code presented on a display) to be readable for the human eye.¹⁷

¹⁴ Although micro formats such as microfilm and microfiche also require dedicated display devices, e.g., microfilm/fiche readers, text stored on these types of carriers do not undergo a binary reformatting. The text is - although at the cost of a certain effort - still readable to the human eye.

¹⁵ See Hillesund (2006) for a discussion on the conditions of text production in different media formats.

¹⁶ See e.g., Gunder (2002) and Hayles (2003).

¹⁷ There are however instances where binary code fulfils both the task of storage and of presentation; e.g., Morse-signals and documents with Braille script inscriptions.

On the process level there is no clear-cut distinction between the text as sign and the text as image. The files, resulting from an image-based capture process, can be managed in an image format (e.g., Jpg) or be OCR-processed and converted into text format (e.g., .doc). Manually captured information, e.g., entered by keying, can conversely be managed either in text- or in a vector-based format.¹⁸ This reformatting also has implications for the distribution of documents as it enables transmission in serial as well as parallel sequences.¹⁹

There does not have to be any apparent visual correspondence between a book situated “in real life” and the same document transferred to digital format, e.g., as a OCR-processed text file. The ontological restructuring that takes place during the digitization process consequently often gives rise to the assumption that text based information can be treated as a phenomenon that exists independently of the document’s characteristics as a material artefact, as if the digitization process would generate “virtual information objects” stripped of all material and contextual characteristics. As demonstrated by Pédaque (2003) this is obviously not the case although some of the rethoric relating to new media seems to suggest the opposite.²⁰

Despite its apparent dematerialisation, digitally formatted information is subject to the same set of prerequisites as a printed book; materiality, textual inscription, and social context, although in this instantiation it is represented by a configuration involving computer hard- and software, interfaces, networks, and the transfer of text based information between states of binary and alpha numeric encoding. The performance of the text is intrinsically dependent on the total informative configuration of the document.

7. Transmissional Alterations

The transmission process is often seen as affecting the information capacity of reproductions in a negative way. The comments by Yeo (2010) are indicative of this: “*But whatever their form, copies and descriptions, like all representations, introduce some loss. ...Every link in the chain adds more distortion....*”²¹ Reproduction is an activity that results in products subjected to a varying degree of alteration in relation to their source documents. But these alterations do not necessary result in losses as information also can be added during the process.²² Such changes can also simultaneously effect different aspects of the reproduction itself.

Alterations can be classified according to intentionality. Examples of *unintended* (or at least unforeseeable) *enrichment* can be found in crowd source activities, e.g., manual OCR correction performed as a computer game²³ or as user driven enriching of catalogue posts and interacting via open APIs.²⁴

¹⁸ See Biggs (2004) for a discussion on the impact of file formats in relation to text- or image based file content.

¹⁹ It has to be noted that what is transmitted is actually the instructions that controls the text’s performance not the actual textual signs.

²⁰ E.g., Lewis in an article aptly named: *What if libraries are artifact-bound institutions?* As bits on the worldwide network, information is freed from the artefacts that have traditionally contained it. Networked information does of course have a tangible form somewhere, and this collection will still require management, but its use does not require a tangible container. (Lewis 1998, 191)

²¹ Yeo (2007, 341)

²² The JISC project “Enriching digital resources programme” JISC (2008) presents several of examples of how the informative capacity of source documents and collections is enhanced by, for example, development of thematic clusters, information extraction, transcription and OCR processing metadata enhancement, improved interaction, creation of user generated content, data sharing, development of pedagogical tools and functionalities.

²³ “*DIGITALKOOT*” the National Library of Finland

²⁴ Powerhouse Museum, Sydney

Unintended loss occurs when the characteristics of the reproduction do not meet the targets prescribed by quality management. *Intended* alterations—*losses* as well as *enrichments*, encompass changes planned and performed within controlled parameters.

There are several instances during the process of transmission where alterations might occur. Dahlström (2004) structures such sources in four categories:²⁵

- the socio-cognitive, psychological, linguistic particulars of the individual(s) responsible for carrying out the translation,
- socio-cultural and socio-technical particulars of the situation in which the translation takes place (e.g., culture and tradition, purpose, specific audience, media environment),
- the material and technological particulars of the departure and target media (such as supporting matter, longevity, compatibility, document architecture),
- physical or symbolic tools at use in the process (such as practices and techniques, software, platforms, requirements, regulations and rules) and so on.
- (Dahlström 2004, 22-23)

Transmissional alteration is in this respect a phenomenon that inevitably will characterise the reproduction and also define its relation to the source document. What kinds of alterations are acceptable? If the digitization process is understood in terms of the conduit model,²⁶ the ambition to filter out alterations—noise—that may appear during the transmission is understandable.

But if we regard digitization as comparable to other strategies for transmission that have been used historically (e.g., manual transcription and letter press printing), we have to consider the alterations introduced in the process in a more inclusive way.²⁷ Pearsall argues in relation to transmission of documents in manuscript format that

...each act of copying was to a large extent an act of recomposition, and not an episode in a process of decomposition from an ideal form.(Pearsall 1984, 127)

All methods of transmission will cause alterations, some of which will become integrated aspects of the resulting surrogates, either momentary, due to the effects of intermediacy²⁸ or as a result of historical processes being integrated as constituent parts of the document, e.g., traces of production processes form manuscript practice such as pricking and ruling. Yeo (2010) observes that the products of reproduction processes such as migration, transcription or conversion, although they always differ in some aspects in relation to the source document, will eventually become originals themselves, when integrated in institutional practice.²⁹

²⁵ The use of “translation” in Dahlström (2004) is equivalent with my use of “transmission.”

²⁶ See note 9.

²⁷ As Levinson observes, noise may even be regarded as representative feature of a specific transmission configuration. We need to acknowledge that the defence of any older communication system may be grounded in a not necessarily illogical desire to hold on to the noise we have in favor of the noise we may get. (Levinson 1997, 52)

²⁸ See Bolter and Gruisin (1991). The remediating effect of digital technology is commented by Deegan and Sutherland in relation to a database dedicated to early printed books: ...the screen image appears as the real thing - to such a degree that the absence of binding, front and end pages, blanks, special front matter (so much of the material that constitutes a book as distinct from its text) goes unnoticed. (Deegan and Sutherland 2009, 133)

²⁹ (Yeo 2010, 112)

8. Structures of Stability

Levy (2000) states

Differences will always be introduced in copying; the trick is to regulate the process sufficiently so that the resulting differences are of little or no consequence and that the properties of greatest consequence are shared (Levy 2000, 26-27)

The reproduction is always effected by alterations. How can the user be sure that the properties of greatest consequence really are shared, that the “*record can stand for the fact it is about*”? This question brings us back to the concept of reliability

There appear to be two principle ways of addressing this issue. One is to focus on the source document and define aspects that appear to be of significance. Concepts such as “*essential characteristics*” (Puglia and Rhodes 2007) “*Key features*” (Chapman and Kenney 1996) have been used in order to describe such criteria given the particulars of the document in question. The reproduction process will consequently be designed to capture and render these aspects. However, if we follow the assumption that the informative capacity of the document is not an intrinsically fixed property, we have to accept that attempts to develop generally applicable specifications regarding any distinct set of informative characteristics will risk omitting other equally informative aspects.³⁰

An alternative approach is to consider the multidimensional information potential of the document and to regard the contextual embeddedness as well as the material and cognitive characteristics of the source document as equally important aspects.

Such an approach corresponds to the views purported by MacNeil and Mak (2007) who use the concept “*modes of stabilisation*” to describe how the institutional context defines the function of documents.

It is indeed the place of librarians and archivists to place a structure of stability over what seems to be an endless flow of infinite possibilities. Some resources will require different modes of stabilization than others. Some resources may require more stabilization than others; this will depend on what the material in question is, and wherein its capacity to generate consequences is located. (MacNeil and Mak 2007, 46)

This process of stabilisation will inevitably highlight some aspects of the source document at the expense of others. Digitization can thus be described as a process whereby a specific informative configuration of a document is stabilised (in this instance by the use of digital technology) in relation to a given function.

Drucker (2008) refers to this flexibility as she states that:

...any artifact (print or digital) is available for interpretation along an infinite, inexhaustible number of lines of inquiry. [...] Words, graphics, images, texts, notes, smudges, marks, white space – in short, any materially present bit of the artifact can serve as a provocation for interpretation. (Drucker 2008, 4)

³⁰ For a thorough discussion on the implications of different approaches to issues of similarity see Yeo (2010).

Similar concerns can be discerned in Hjørland's (1998) use of the term "*epistemological potentialities*"³¹ of a document. Duranti (2003) observes that an archival record has evidentiary capacity to support an "*infinite number of facts through its content, its form, its archival bond, and its context.*"³² Also within museology, we find references to the object as "*...a limitless source of information.*"³³

Digitization as a process should focus on the twofold characteristics of a document. The analysis requires a theoretical framework that assigns a certain amount of consistency to the document in order for a specific informative configuration to be persistent throughout the process. However, at the same time, this framework has to guarantee enough flexibility to accommodate the functional differences between the analogue and the digital domain and to avoid the static limitations of the channel metaphor.

The "*capacity to generate consequences*"³⁴ is determined by the relation between the characteristics of the document and the mode of stabilisation used—the particulars of the digitization process. The question of reliability of a digital reproduction could then be viewed as the capacity of the process to transfer and re-establish a specific informative configuration—from analogue to digital format—while allowing for the "*completeness of the record's form and the amount of control exercised on the process of its creation*"³⁵ to be examined.

9. Conclusion

The information potential of the source document remains open to interpretation until its informative configuration is stabilised, in accordance with its intended function. This informative configuration is represented as a new document in digital format, based on the parameters of the digitization process, a document that subsequently is exposed to new sets of interpretation either through direct use or via further remediation processes.

The digitization process always "leaves its mark" on the product—the reproduction.³⁶ The alterations that are generated represent an inevitable and intrinsic contribution to any transmission.³⁷ Digitization can in this respect be described as an act of interpretation. This digital malleability also results in a de-contextualisation as we interact with various *representations* of the documents. The interpretative capacity can be used to re-establish the informative configuration of the source document; it also can be used to enhance or emphasise specific aspects of the source document and its contextual setting.

The question of reliability is consequently to a large extent dependent on whether the user gains insight into the convergence and transformation processes inherent in the digitization process. Differences in institutional practice reflect the way in which the information capacity of the objects in their collections is defined. We find that libraries tend to view documents primarily as providers of text based

³¹ Hjørland (1998, 610)

³² Duranti (2003, 8)

³³ van Mensch (1992)

³⁴ MacNeil and Mak (2007, 46)

³⁵ InterPARES 2 Project

³⁶ Dahlström observes, in the translation process, certain features of the work are preserved that can be carved into the flesh of the new medium and be expressed by its architecture and the language of its web of signs, while others are treated as noise, obscuring the substantial signals. (Dahlström 2004, 23)

³⁷ A parallel can be found with other institutionalised processes of transmission such as scholarly editing and conservation. See Eggert (2009).

information.³⁸ As demonstrated by the R.T. Pédaque framework, text based information content is dependent on the interplay between all three informative dimensions for the document to fulfil its informative function. An uncritical focus on one isolated aspect, i.e., the textual inscriptions, will reduce this informative capacity.

The concept of informative configurations can be used to exemplify how different interests influence the design of digitization processes and consequently also the characteristics of the resulting products. As the full informative potential of the document cannot be transmitted (unless we were able to produce clones), the user's ability to evaluate the informative capacity of the reproduction becomes ever more important.

The concept of "reliability", as defined by the InterPARES framework demonstrates that this ability is not a fixed property, but a matter of establishing a condition of trust; the user of a reproduction has to accept the propositions made by its producer.

References

- Abram, Stephen, and Judy Luther. "Born with the Chip - 05/01/2004 - Library Journal." *Library Journal.com*. 2004.
- Biggs, Michael A.R. "What Characterizes Pictures and Text?" *Literary and Linguistic Computing* 19, no. 3 (2004): 265–272.
- Bolter, Jay David, and Richard Grusin. *Remediation : Understanding New Media*. Cambridge, Mass.: MIT Press, 1999.
- Brown, John Seely, and Paul Duguid. "The Social Life of Documents." *First Monday* 1, no. 1-6 (1996).
- Caple, Chris. *Conservation Skills: Judgement, Method and Decision Making*. London: Routledge, 2000.
- Chapman, Stephen, and Anne R. Kenney. "Digital Conversion of Library Research Materials." *D-Lib Magazine* (October 1996).
- Conway, Paul. "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas." *Library Quarterly* 80, no. 1 (2010): 61–79.
- Cornell University. "Moving Theory into Practice - Digital Imaging Tutorial." http://www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html.
- Dahlström, Mats. "How Reproductive Is a Scholarly Edition?" *Literary and Linguistic Computing* 19, no. 1 (April 2004): 17–33.
- Dahlström, Mats. *Under Utgivning : Den Vetenskapliga Utgivningens Bibliografiska Funktion*. Skrifter Från Valfrid, 1103-6990 ; 34. Borås: Valfrid, 2006.
- Deegan, Marilyn, and Kathryn Sutherland. *Transferred Illusions: Digital Technology and the Forms of Print*. Farnham: Ashgate, 2009.
- Di Leo, Jeffrey R. "The Cult of the Book-and Why It Must End." *The Chronicle Review* (September 26, 2010). <http://chronicle.com/article/From-Book-to-Byte/124566/>.

³⁸ See Maroevic (1998) for a study of how definitions of informative capacity of objects and documents vary within the MLA sector according to institutional type.

- Drucker, Johanna. "Developing Interpretation in E-Space: Humanities 'Tools' in Digital Environments 3." *Humanities Tools Part 3*. UCLA, November 17, 2008.
http://polaris.gseis.ucla.edu/drucker/humanitiestools_pt3_int.pdf.
- Duranti, Luciana. "More Than Information, Other Than Knowledge: The Nature of Archives in the Digital Era." *Cadernos BAD* (2) (2003): 6–16.
- Eggert, Paul. *Securing the Past : Conservation in Art, Architecture and Literature*. Cambridge, UK: Cambridge University Press, 2009.
- FADGI, Federal Agencies Digitization Initiative - Still Image Working Group. *Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files*. 2010.
http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf.
- Francke, Helena. "What's in a Name? Contextualizing the Document Concept." *Literary and Linguistic Computing* 20, no. 1 (2005): 61–69.
- Furner, Jonathan. "Information Studies Without Information." *Library Trends* 52, no. 3 (2004): 427–446.
- Gunder, Anna. "Forming the Text, Performing the Work: Aspects of Media, Navigation, and Linking." *HumanIT* 2-3 (October 24, 2002).
- Hillesund, Terje. "Digital Text Cycles: From Medieval Manuscripts to Modern Markup." *Journal of Digital Information* 6, no. 1 (February 2006).
- Hjørland, Birger. "Theory and Metatheory of Information Science: a New Interpretation." *Journal of Documentation* 54, no. 5 (1998): 606–621.
- Hjørland, Birger. "Documents, Memory Institutions and Information Science." *Journal of Documentation* 56, no. 1 (2000): 27–41.
- IFLA. "Guidelines for Digitization Projects for Collections and Holdings in the Public Domain." 2002.
<http://www.ifla.org/en/publications/guidelines-for-digitization-projects-for-collections-and-holdings-in-the-public-domain>.
- InterPARES 2 Project. *Terminology Database*. http://www.interpares.org/ip2/ip2_terminology_db.cfm.
- JISC. "Enriching Digital Resources 2008-09." 2008.
<http://www.jisc.ac.uk/whatwedo/programmes/digitisation/enrichingdigi.aspx>.
- Latour, Bruno. "Visualisation and Cognition: Drawing Things Together." *Knowledge and Society: Studies in the Sociology of Culture and Present* 6 (1986): 1-40. <http://www.bruno-latour.fr/>.
- Levinson, Paul. *The Soft Edge: a Natural History and Future of the Information Revolution*. London: Routledge, 1997.
- Lewis, David W. "What If Libraries Are Artifact-Bound Institutions?" *Information Technology and Libraries* 17, no. 4 (1998): 191–197.
- Levy, David M. "Where's Waldo? Reflections on Copies and Authenticity in a Digital Environment." In *Authenticity in a Digital Environment*, 24–31. Washington D.C.: Council on Library and Information Resources, 2000.
- Levy, David M. "Documents and Libraries: A Sociotechnical Perspective." In *Digital Library Use: Social Practice in Design and Evaluation*, edited by Ann P. Bishop, Nancy A. Van House, and Barbara Pfeil Buttenfield, 25-42. MIT Press, 2003.
- Lund, Niels Windfeld. "Document, Text and Medium: Concepts, Theories and Disciplines." *Journal of Documentation* 66, no. 5 (2010): 734–749.

- MacNeil, Heather, and Bonnie Mak. "Constructions of Authenticity." *Library Trends* 56, no. 1 (2007): 26–52.
- Maroevic, Ivo. "The Phenomenon of Cultural Heritage and the Definition of a Unit of Material." *Nordisk Museologi* 2 (1998): 135–142.
- van Mensch, Peter. "Towards a Methodology of Museology," Ph.D. Thesis, University of Zagreb, 1992.
http://www.muuseum.ee/et/erialane_areng/museoloogiaalane_ki/ingliskeelne_kirjand/p_van_mensch_towar/.
- Meyrowitz, Joshua. "Multiple Media Literacies." *Journal of Communication* 48, no. 1 (1998): 96–108.
- Muñoz Viñas, Salvador. *Contemporary Theory of Conservation*. Oxford: Elsevier Butterworth-Heinemann, 2005.
- National Library of Finland. "Digitalkoot." *Digitalkoot*. <http://www.digitalkoot.fi/en>.
- Pearsall, Derek. "Texts, Textual Criticism, and Fifteenth Century Manuscript Production." In *Fifteenth-century Studies: Recent Essays*, edited by Robert F. Yeager, 121–136. (Archon Books, 1984..
- Pédaque, Roger T. "Document: Form, Sign and Medium, As Reformulated for Electronic Documents." ver. 3, July 8, 2003.
http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/05/11/index_fr.html.
- Powerhouse Museum, Sydney. <http://www.programmableweb.com/api/powerhouse-museum-collection>.
- Puglia, Steven, and Erin Rhodes. "Digital Imaging - How Far Have We Come and What Still Needs to Be Done?" *RLG DigiNews* 11, no. 1 (April 2007).
- Ranganathan, S. R.. *Documentation and its facets; being a symposium of seventy papers by thirty-two authors*. Bombay; New York: Asia Pub. House, 1963.
- Ross, Seamus. "Progress from National Initiatives Towards European Strategies for Digitisation." In *Towards a Continuum of Digital Heritage: Strategies for a European Area of Digital Cultural Resources, European Conference*, 88–98. Den Haag: Dutch Ministry of Education, Culture and Science, 2004.
- Smiraglia, Richard P. *The Nature of 'a Work' : Implications for the Organization of Knowledge*. Lanham, Md.: Scarecrow, 2001.
- Shannon, C. E. "A Mathematical Theory of Communication." Reprinted with Corrections from *The Bell System Technical Journal* 27(July, October, 1948): 379–423, 623–656. <http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>.
- Talja, Sanna, Kimmo Tuominen, and Reijo Savolainen. "'Isms' in Information Science: Constructivism, Collectivism and Constructionism." *Journal of Documentation* 61, no. 1 (2005): 79–101.
- Warwick, Claire, Isabel Galina, Jon Rimmer, Terras Melissa, Ann Blandford, Jeremy Gow, and George Buchanan. "Documentation and the Users of Digital Resources in the Humanities." *Journal of Documentation* 65, no. 1 (January 2009): 33–57.
- Xie, Sherry. "Building Foundations for Digital Records Forensics: A Comparative Study of the Concept of Reproduction in Digital Records Management and Digital Forensics." *American Archivist* 74, no. 2 (September 2011): 576–599.
- Yeo, Geoffrey. "Concepts of Record (1): Evidence, Information, and Persistent Representations." *The American Archivist* 70 (2007): 315–343.
- Yeo, Geoffrey. "'Nothing Is the Same as Something Else': Significant Properties and Notions of Identity and Originality." *Archival Science* 10 (2010): 85–116.

Preservation of Audiovisual Material

The Media Preservation Initiative at Indiana University Bloomington

Mike Casey

Indiana University

Abstract

In 2009, Indiana University Bloomington published a report documenting the findings of a campus-wide census of audio, video, and film holdings which identified more than 560,000 media objects, most of them on degrading, obsolete analogue carriers. In 2010, Indiana University began a preservation planning process for time-based media holdings, engaging key campus stakeholders. The results of the first year of this planning project were published in 2011 in a document entitled Meeting the Challenge of Media Preservation: Strategies and Solutions. This paper explores the central components of these publications including key recommendations of the Indiana University Media Preservation Initiative to create a centralized digitization facility and a campus-wide media preservation plan. To our knowledge, no other university in the US has produced a comprehensive, in-depth plan to preserve media holdings distributed across a campus. This may provide a model or stimulate the thinking of others with similar problems and needs.

Author

Mike Casey is the Director of Media Preservation Services for Indiana University's Media Preservation Initiative. He is the co-author of Sound Directions: Best Practices for Audio Preservation, a contributing author to the second edition of the best practices publication IASA-TC 04, and the creator of FACET: The Field Audio Collection Evaluation Tool. He also authored the Indiana University Media Preservation Survey report and is the principal author for the follow-up report Meeting the Challenge of Media Preservation: Strategies and Solutions. He is adjunct faculty in the School of Library and Information Science where he teaches a class in audio preservation.

"...in the mid- to long-term there is a major risk that carrier degradation combined with playback obsolescence will defeat the efforts of archivists to ensure the survival of the content in their care."

--International Association of Sound and Audiovisual Archives¹

1. Introduction

Fifteen years is a short period of time in the preservation world where work proceeds for generations and time horizons are very long. Yet, many media preservation practitioners believe that there is only a 15 year window of opportunity to digitally preserve audio and video recordings due to active degradation and rapidly expanding obsolescence. After that, many think that preservation transfer will be impossible, achievable only with diminished fidelity, or prohibitively expensive, particularly for large collections.

¹ Task Force to establish Selection Criteria of Analogue and Digital Audio Contents for Transfer to Data Formats for Preservation Purposes, ed. Majella Breen et al. (Hungary: International Association of Sound and Audiovisual Archives, IASA Editorial Group), accessed August 28, 2012, <http://www.iasa-web.org/task-force/2-introduction>.

With this in mind, the Indiana University (IU) Archives of Traditional Music (located in Bloomington, Indiana in the U.S.) ran a few numbers several years ago to answer a critical question: how long would it take its one grant-funded audio engineer, transferring one recording at a time, to complete preservation of the Archives' holdings? The answer: 58 years. About the same time, the Indiana University Cook Music Library performed this same exercise and calculated an answer of 120 years for their holdings.

We quickly recognized that we had a serious problem on our campus that demanded a larger solution implemented at a higher level. From this realization the IU Bloomington Media Preservation Initiative (MPI) was born. From the beginning, MPI drew upon existing campus experience and past projects concerned with media preservation issues. For example, the NEH-funded *Sound Directions* project, a collaboration between Indiana University and Harvard University that resulted in an internationally-used best practices publication for audio preservation as well as software tools including the Field Audio Collection Evaluation Tool (FACET) and the Audio Technical Metadata Collector (ATMC). (see <http://www.dlib.indiana.edu/projects/sounddirections/>) Another example is the *EVIA (Ethnographic Video for Instruction and Analysis) Digital Archive Project* (<http://www.eviada.org/>) that developed software and systems for the annotation, discovery, peer review, and scholarly publication of video as well as created a digital archive of ethnographic field video for use by scholars and instructors. These and other projects provided a baseline of experience from which to begin work.

2. Media Preservation Census

The first task for the MPI was to define problems and challenges. We began with a preservation census of campus holdings, the results of which were published in 2009. This document, titled *Media Preservation Survey: A Report*, is available from <http://www.indiana.edu/~medpres/index.shtml> In addition to presenting data on holdings, the report explores degradation and obsolescence issues found in campus collections as well as media recordings in general. The report also includes chapters on the research value of holdings, physical storage conditions, reformatting efforts, discovery and use, and existing campus resources.

The broad outlines of what we learned about our holdings can be summarized as follows:

- There are more than 560,000 audio, video, and film objects on the Bloomington campus
- 64% are audio, 22% video, 14% film
- Located in more than 80 campus units
- Held on over 50 formats
- Estimated 44% are unique or rare
- Dating from 1893-present
- Large numbers have national or international value

Essentially, we learned that our problem was not only serious but both larger and wider in scope than anticipated. The Bloomington campus of Indiana University holds very large numbers of media objects that are actively degrading, carried on formats that are obsolete, considered highly valuable for research and instruction, for which we have a relatively short time window to take preservation action. Accordingly, we needed strategies and solutions that could be employed both rapidly and on a massive scale, yet engaged preservation standards and best practices. We strongly suspect that other archives and holders of media recordings are faced with a similar problem and have the same needs.

3. Planning Process and Report

The survey report led to a year-long preservation and access planning project guided by a task force appointed by the IU Bloomington provost. This project was funded by key campus administrative units: the Office of the Provost, Office of the Vice Provost for Research, IU Libraries, University Information Technology Services, and The College of Arts and Sciences. We established the following structure to undertake this project:

- Working Group consisting of six members charged with day to day and week to week research for the project. This group met once a week
- Task Force consisting of 10 members (including the members of the Working Group) charged as the official body to make recommendations. This group met every six weeks or so
- IU Bloomington Advisory Board consisting of key campus stakeholders charged with advising the Task Force. This group was convened twice during the project period
- External Advisory Board consisting of international media preservation experts convened once during the 2010 IASA/AMIA conference
- Consultant, AudioVisual Preservation Solutions, to assist with all parts of the project including data projections, building plans, etc.

Completed in 2011, this planning process resulted in a publication titled *Meeting the Challenge of Media Preservation: Strategies and Solutions* that is available from <http://www.indiana.edu/~medpres/index.shtml>

4. Indiana Media Preservation and Access Center

Meeting the Challenge charts solutions and lays the groundwork for unlocking campus media assets and transforming them into usable resources. Perhaps the cornerstone recommendation is to build the Indiana Media Preservation and Access Center (IMPAC), a service unit that will provide audio and video preservation transfer as well as digitization of film for campus holdings. This recommendation is built upon an analysis of the advantages and disadvantages of undertaking the digitization stage in-house versus outsourcing. This analysis examined variables such as scale and uniformity of holdings, quality control, desired future location of expertise, potential for supporting the academic mission of the University by providing educational services, national leadership opportunities, and cost. It also relied upon the realization that digitization is but one of many steps in the preservation process. IU Bloomington must develop preservation infrastructure if it is to be successful, regardless of where digitization is completed.

The IMPAC will provide the services necessary to attain campus targets, which include the preservation of 284,000 audio recordings and 66,000 video recordings along with access digitization of 58,000 films, all within 15 years. These services will include:

- preservation transfer (digitization) of analogue audio recordings;
- preservation transfer of analogue video recordings;
- preservation transfer of physical digital (CD, MiniDv, DVD, for example) audio and video recordings;

- digitization of motion picture film;
- photographs of preserved objects and their containers;
- digitization work to fulfill orders from researchers for campus media holdings;
- creation of derivative digital files for researcher access;
- collection of technical and digital provenance metadata on the preserved object, resulting digital files, and the preservation process;
- preparation of media holdings for digitization;
- assistance with prioritization of media holdings for preservation treatment.

Meeting the Challenge presents a detailed build plan for the IMPAC, estimating number/type of staff, number/type of media studios, facility square footage, and digital storage over the life of the project. This build plan was developed through a strongly data-driven process. Our recommendations were derived directly from data on campus time-based media collections combined with analysis by MPI with assistance from consultant AVPS. We worked from the inside out, letting the data lead us to conclusions regarding the size and scope that are necessary to reach our preservation targets within the 15 year time frame.

The IMPAC approach to preservation transfer addresses preservation concerns while utilizing higher throughput workflows, placing a strong emphasis on maintaining preservation principles within a high efficiency approach. Choosing transfer workflows and approaches forced us to define where the intersection of preservation principles and efficiency lies for our institution. While parallel transfer workflows (simultaneous digitization of multiple recordings) may result in larger numbers of items digitized per unit of time, they also carry a higher risk that the products of digitization are not optimal. Our plan is to use a mix of smaller scale parallel transfer workflows (expected to be 4:1 and 2:1) along with custom 1:1 work when necessary. The parallel transfer workflows, for which we will employ risk mitigation procedures to address preservation issues, will greatly increase the output of products. Some use of parallel transfer workflows is necessary given the large numbers of recordings that must be preserved within a short time frame.

5. Prioritization

The report also recommends a prioritization process for campus holdings that utilizes both software applications and curatorial expertise. It is clear that not every recording can or will be preserved in time due to degradation and obsolescence issues as well as the large numbers of items held in Bloomington. It is also true that there are no guarantees—for example, economic slowdowns and recessions cannot be predicted and can result in loss of resources. Plus, not every recording or collection is considered of equal value. We feel that it is a wise strategy to determine which of our holdings represent our institution's highest priorities and to get these preserved as soon as possible.

The prioritization process relies upon a combination of software tools and curatorial expertise to assess preservation condition/risk and research/instructional value. The tools assist with structuring the analysis and provide a measure of objectivity as well as transparency to what is unavoidably, in part, subjective work. However, we do not believe that prioritization decisions can be left to software applications alone. Many parts of this process are guided by the expertise and experience of unit curators and collection managers as well as media technical and format experts. MPI staff provide technical expertise for analysing risk, preservation condition, and obsolescence while unit staff provide content expertise for analysing research and instructional value. Here is an overview of the process:

5.1 Meeting with MPI team and unit staff

One purpose of this meeting is to select collections or other groupings of media recordings to evaluate during this stage of prioritization. Since our immediate goal is a five-year plan, this is just the first of several rounds of prioritization, and we will not be able to evaluate all holdings. We will rely upon unit staff to identify high-value collections. We will also assist in determining which collections are most at risk or in the poorest condition.

5.2 Analysis of risk, preservation condition, and obsolescence

The MPI team will use a software application to score collections for risk, condition, and obsolescence. This will involve gathering and analysing data from a visual inspection of the collections under consideration.

5.3 Analysis of research and instructional value

The MPI team will assist curators and collection managers in using a software application to score collections for research and instructional value. The MPI team will also help units research their collections, gathering data as needed to feed into this process. This step will be driven by unit curators based on their judgment of the value of their holdings.

5.4 Curatorial review

In this step, curators and/or collection managers will examine the rankings of their collections and make adjustments as necessary. They may also take into account other considerations that impact value including such things as timeliness (upcoming events or anniversaries), publicity opportunities, and others.

5.5 Validity

The final rankings for any given unit will be valid within the context of that unit only. It is difficult to rank collections with consistency and integrity across units, not to mention reaching agreement across campus on the relative value of the various and diverse media collections. For these reasons, we will try to achieve consistent rankings within each unit only. MPI will then highlight each unit's top priorities as campus preservation priorities. This enables unit curatorial staff to maintain significant control over the prioritization process for their content.

The Indiana Media Preservation and Access Center, when built, will have the capacity and operating efficiency to guarantee that top priorities will be preserved within the defined 15-year preservation period.

To support this process, MPI has developed a software tool to assist the assessment of research and instructional value of media collections. We have also developed another selection for preservation tool to evaluate preservation condition, risk, and obsolescence of audio, video, and film collections. This tool leverages previous software development work at several other academic institutions. Both of these applications will be made freely available in 2013.

6. Preservation Principles

The foundation for all of the recommendations discussed in *Meeting the Challenge* is a set of preservation principles articulated in the report. These principles guide the development and implementation of preservation strategies so that efficient, accurate, sustainable, and enduring work is supported as well as cooperation between stakeholders, all while maintaining a consistent focus on the primary goal of long-term preservation. The principles are presented in full form in the report but may be summarized in encapsulated form as follows:

- A long time horizon must frame all decisions
- Timely decisions must be made to combat obsolescence and degradation
- Digitize once—it will not be feasible a second time
- Preservation digital files must reflect the concepts of faithful reproduction, accuracy, and integrity
- Ongoing preservation is required, not just one-time digitization
- Use international standards and best practices where they exist
- Preservation and access must not compromise each other
- Leverage existing IU resources
- Build strong partnerships
- Long-term preservation, access, and management decisions reside with curatorial and unit staff
- Transparency to stakeholders
- Prioritization before preservation
- High-efficiency workflows are needed to meet targets

7. Access Principles

Of course, the holdings of Indiana University are preserved so that they may be accessible and provide for a variety of uses into the foreseeable future. Access, broadly defined, includes the discoverability, the deliverability, and the usability of any given media item. Access to preserved holdings is critical to the success of the IMPAC and to the realization of its value to the campus. *Meeting the Challenge* presents a set of guiding access principles to serve as the foundation for the development of specific access policies and procedures. These principles are articulated in a number of broad areas including curatorial responsibility, standards and best practices, online accessibility, infrastructure support, description and cataloging services, metadata, copyright strategies, rights management tools, timeliness of delivery, derivative quality and management, and others. An access working group is currently developing specific policies and procedures based on these general principles.

8. Campus Engagement

Perhaps none of our recommendations will be adopted if we do not engage Indiana University Bloomington's research, teaching, and service missions. The report analyses how our strategies build upon existing campus resources and strengths, engage campus strategic plans, contribute to the university's instructional mission, and provide clear opportunities for national leadership. In other words, we not only have to solve the media preservation and access problem, we must also demonstrate that our efforts will provide solid contributions to our institution. Of course, the products of MPI work—preserved and accessible media collections—will transform research and instruction for faculty and students whose

work can benefit from media resources. However, we believe that it is also critical to show how this work will contribute to the advancement of Indiana University in as many ways as possible.

In addition, *Meeting the Challenge* includes chapters exploring strategies for film, technology infrastructure needs, and facility planning including a general build plan. This report and the earlier media preservation survey report, as well as the process we used to gather data and plan, may be useful to other institutions addressing the long-term preservation and access of their media holdings.

9. Ongoing Work

The work of the MPI continued during fiscal year 2011-12 and is continuing during fiscal year 2012-13 as well. One major new objective is to leverage existing campus resources to begin audio and video preservation work plus film conservation before the IMPAC is constructed. We call this the IMPAC startup project. In effect, the work of the IMPAC is starting before the Center is established or has a central physical location. It is slow but steady, limited but effective and includes:

- Audio preservation transfer work for a number of campus units based at the Archives of Traditional Music. This utilizes former *Sound Directions* staff, workflows, and infrastructure. It also includes contributions from the Music Library and School of Music staff;
- Video preservation pilot projects and transfer work at Radio & Television Services utilizing their engineering staff;
- Work on film collections under the guidance of the IU Libraries film archivist at the Auxiliary Library Facility.

Startup project objectives are to test proposed workflows, demonstrate proof of concept, and gain further experience while creating a small body of high-value preserved/conserved content that we can point to with stakeholders and potential donors. And, of course, we will preserve a few key campus holdings. While we must build the IMPAC to make real progress in preserving campus media holdings in time, the startup project lays the foundation and moves us a step or two forward.

To aid in building and operating the IMPAC as well as increasing available resources, MPI is currently exploring possibilities for partnerships. This may take the form of a public-private partnership with a private media preservation company and/or a public-public partnership with one or more academic institutions. Partnerships such as these can bring greater experience and expertise to the project as well as additional resources and content. It is clear to us that successful long-term preservation of very large media holdings such as those found not only at IU but at other institutions will require cooperation, collaboration, and strong partnerships in many ways.

In the past year MPI has also developed a communications strategy, including a blog where campus stakeholders and interested parties from outside institutions can follow our progress. The blog may be accessed at <http://mediapreservation.wordpress.com/>.

While we have made significant progress over the past few years, there is much work ahead of us before we can claim success. Ultimately, our overall goal is to move forward into a new era of preservation and access for media holdings, an era characterized by a wealth of enduringly preserved and easily accessed media content integrated into campus research and instruction.

Video Compression...For Dummies?

George Blood

Abstract

Over time video formats and carriers become obsolete rapidly while file-based digital video formats are evolving rapidly. The physical carriers of these files are unreliable; manufacturers assume they will be used for acquisition then transferred to hard disc drive for production. For these reasons our starting assumption is that all historic video formats must be migrated to the latest digital technology. Through this paper, we discuss, in detail, standards and recommendations regarding video compression.

Author

George Blood graduated from the University of Chicago (1983) with a Bachelor of Arts in Music Theory. Actively recording live concerts (from student recitals to opera and major symphony orchestras), since 1982, he has documented over 4,000 live events. From 1984 through 1989 he was a producer at WFMT-FM, and has recorded and edited some 600 nationally syndicated radio programs, mostly of The Philadelphia Orchestra. He has recorded or produced over 200 CDs, 3 of which were nominated for Grammy Awards. Each month George Blood Audio and Video digitizes approximately 2,000 hours of audio and video collections. He is the only student of Canadian pianist Marc-André Hamelin.

1. Background

Two years ago we received a five-year contract from the Library of Congress to digitize audio and video.¹ The Library's in-house standard for video preservation masters is lossless JPEG2000 wrapped in MXF. However, within the Library, only the Culpeper facility has the technology to work directly with this format. As part of this contract, I was asked to prepare a white paper entitled "Determining Suitable Digital Video Formats for Medium-term Storage." In the paper we make recommendations on target formats for video preservation when J2K/MXF is not yet a viable option. Originally, the white paper was intended for use by other departments within the Library of Congress. However, we quickly realized the information and recommendations would be useful to other institutions as well.

Our four starting premises are listed as follows:

- | | |
|--------------------------|---------------------------------|
| 1. Tape is Not an Option | 3. Compression is Not an Option |
| 2. 10-bits Required | 4. One Size Does Not Fit All |

2. The Problem with Tape

Tape was rejected due to obsolescence. Standard definition machines are no longer manufactured, either for analogue or digital formats. Current workflows are rapidly evolving around non-tape, file-based systems, from acquisition to production to distribution.

¹ This paper was originally presented at The Memory of the World in the Digital Age: Digitization and Preservation 26 to 28 September 2012 Vancouver, British Columbia, Canada. We would like to thank the session chairs, Luciana Duranti and Jonas Palm for the opportunity to publish our presentation.

In this paper, I discuss the two middle issues, 10-bit resolution and why compression is not acceptable for preservation, in detail. The observation that one size does not fit all will then form the basis for the structure of the recommendations made in the white paper for the Library of Congress.

3. Requirement of 10-bit resolution

The requirement for 10-bit resolution is the subject of considerable discussion. I begin, therefore, by reviewing how bit-depth works in video. The choice between 8- and 10-bits can be thought of as a sort of compression as it excludes low level detail and softens the image. Let us explore the argument to “use a lower bit rate for lower quality formats.” To appreciate why it is necessary to use 10-bits, let’s explore how this works for audio.

In audio, each bit is equal to 6 decibels (dB). There is a maximum signal level, full scale. As you increase the quantity of bits, you achieve a more dynamic range, or signal-to-noise ratio. Dynamic range and signal-to-noise ratio are simply different ways of looking at the same phenomenon, the range from minimum to maximum information captured.

As you add bits, the dynamic range of information that you can capture also increases. In an 8-bit system you can capture 48dB of dynamic range. In a 16-bit system you can capture 96dB of dynamic range and, finally, in a 24-bit system you have the ability to capture 144dB (Figure. 1)

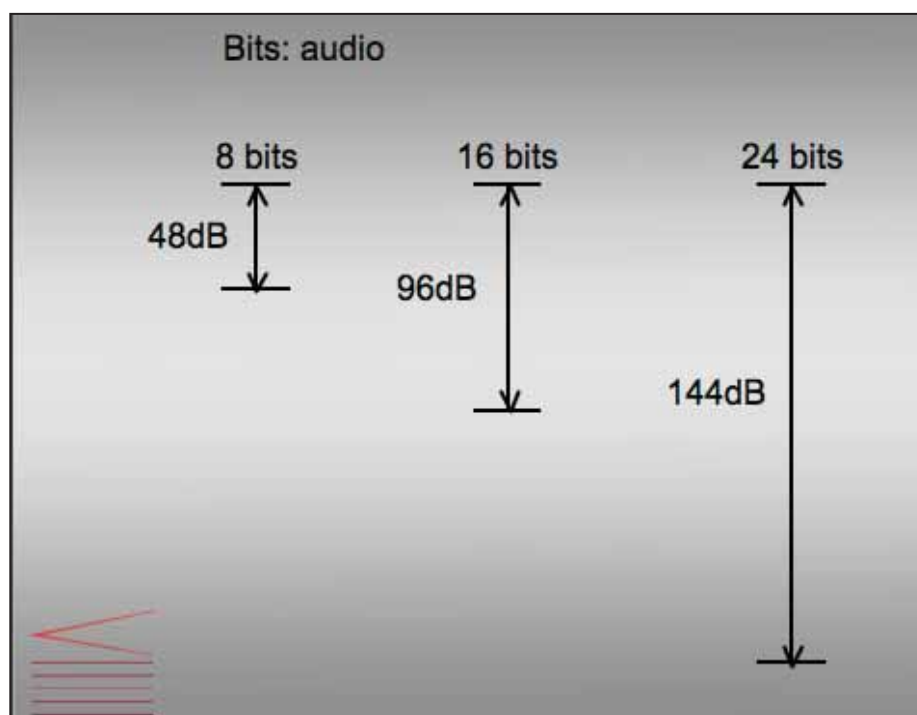


Figure 1. Dynamic range in digital audio is a function of the number of bits in the system.

If your source is, for example, an audiocassette with approximately 58dB of dynamic range, an 8-bit system with 48dB of dynamic range will not be enough. A 16-bit system, such as that used on a compact disc or DAT tape, however, will be more than enough and 24-bits will be wasted storage. However, if

your source is an extremely high quality studio master recording, then the additional resolution and additional dynamic range of a 24-bit system makes sense (Figures 2 and 3).

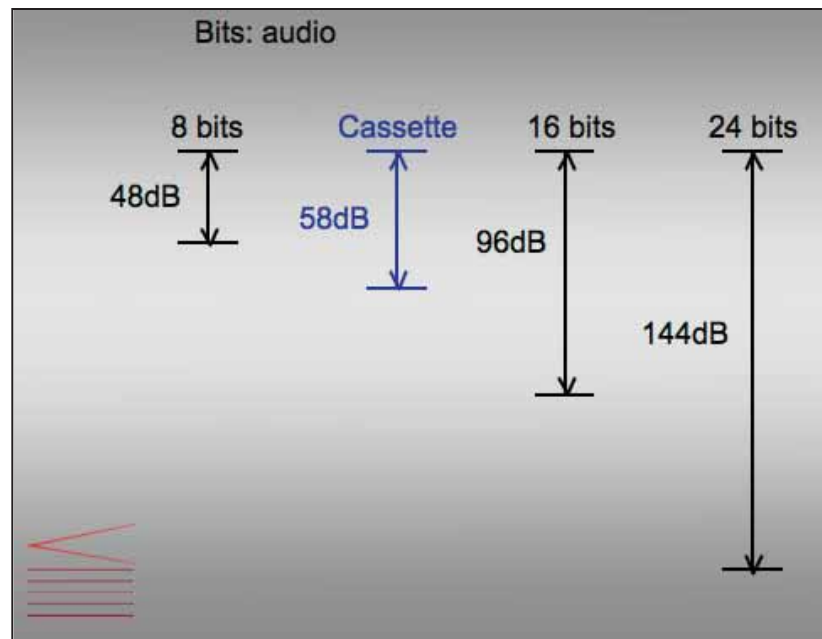


Figure 2. Match dynamic range of system with dynamic range of source media.

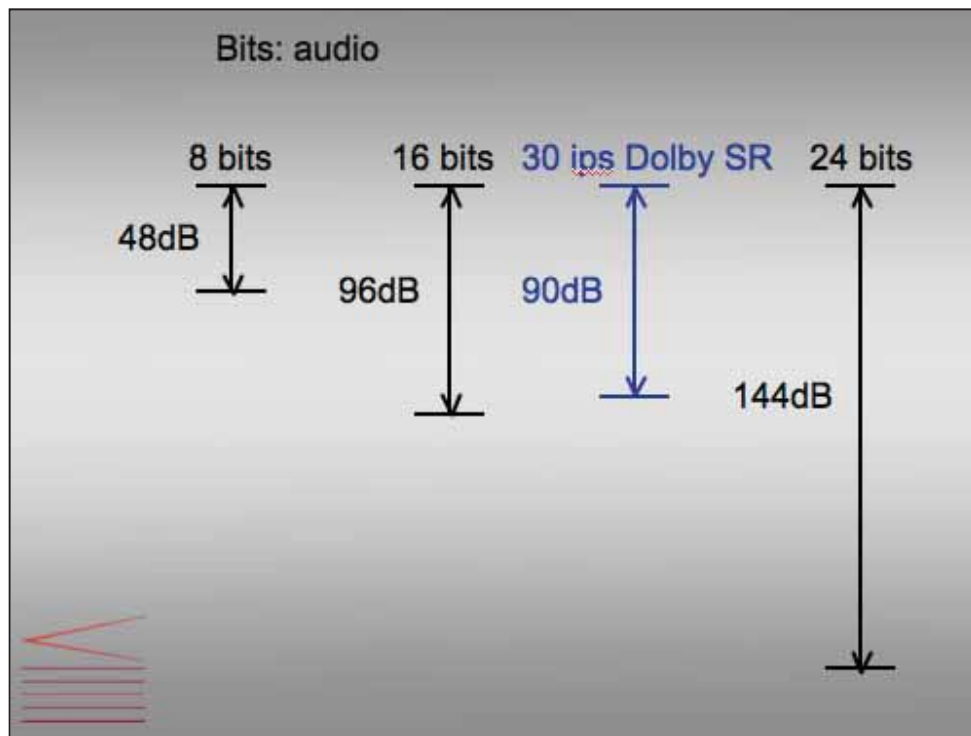


Figure 3. Sources with higher dynamic range (higher quality) need more bits.

In practice, audio is now nearly always digitized at 24-bits for the sake of standardization. In audio, an increased number of bits allows for a wider range of information to be captured. You match the range of the source to the number of bits necessary to capture that range. Storage for audio has also become so inexpensive that it is not cost prohibitive to store that much data. A 1-Terabyte hard-drive can hold up to 500 hours of preservation quality audio and only costs approximately \$100.00. Such high-quality storage for so little expense could certainly not have been achieved with 1/4" tape.

Unfortunately, video does not work in the same way as audio.

The waveform monitor is a tool used to adjust video to technical standards (Figure 4).

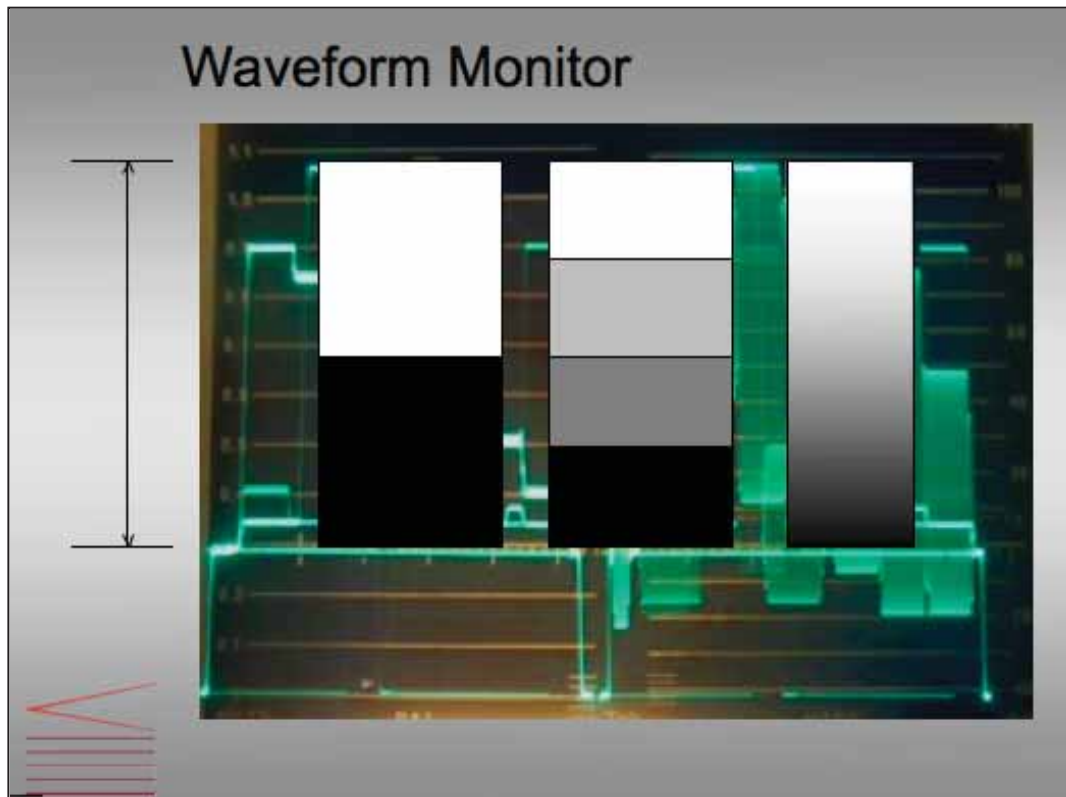


Figure 4. Waveform monitor showing full range of values from 7.5 to 110 IRE. This is colour bars.

A properly recorded analogue video will have voltages in the range between 7.5 and 110 IRE.² At the lower end of the range, it is black and at the other end it is white. If we were to use 1-bit encoding, we would only have these two choices: black and white. As we add bits, we gain more gradations. The goal is a contiguous, smooth transition from black to white. In order to achieve that, a high number of bits are required. Fundamentally the difference between audio and video is that in audio the step size is fixed and the range changes, whereas in video the range is fixed at the step size changes.

² IRE stands for Institute of Radio Engineers. The actual electronic values don't matter for this discussion. The point is there's a defined range and that range doesn't change with the bit depth.

All properly recorded video will contain video levels between 7.5 and 110 IRE, and all of the values in between. This includes VHS and U-matic formats. If they are able to capture black and able to capture white and the source is analogue—analogue being a contiguous signal—the video will contain everything in between. In formats that use colour-under, such as VHS and U-matic, the fewer lines of colour resolution in the vertical does not affect the fact that at each point in time, the entire range of luminance and colors are always available. Indeed, the analogue compression that give us 240 lines of colour resolution makes it significantly more important to be sure to capture the full range of detail in each of those lines.³

In a one bit system half the values would be black and half would be white, just like bi-tonal text scanning. In a two bit system you get black and white plus two shades of gray. As you add more bits, you get finer and finer gradations until you get a very smooth transition from full black to full white. Using fewer bit, even 8, leads to banding, or visible steps. This example is in the luminance channel. The same goes for the two chrominance channels that carry the colour information.

4. Lossy Compression⁴

In an ideal world, lossy compression of video would not be a topic of discussion. We do not accept compression in the preservation of any other format, however some feel that it is acceptable for video.

Using a lower bit rate for lower quality sources sounds like a good idea, but like bit depth in audio, video encoding does not work this way. Once again, let's detour to audio digitization to understand this concept.

DATA RATE AND DIGITIZING AUDIO

A sound wave, in its simplest form, is a sine wave (Figure 5).

Pulse Code Modulation is used in digitizing audio for preservation. This method captures the level of signal at a regular interval in time. The interval is determined by the Nyquist formula, which states that the highest frequency available for capture is one half the sample rate (see Figure 6). A telephone call, for example, has less information than a ½" stereo album master.

If the sample rate is too low for the source signal, information is lost. In Figure 7 there are two signals. The first is a repeat of the one above. The signal is being sampled sufficiently often, according to Nyquist, to capture all the frequency information. In the lower signal there is information between the samples that is not encoded and will be lost. In the lower waveform, the sample rate, the data rate, is not high enough to capture the information in the signal.

To capture all the information in the lower signal, more samples must be taken, the data rate increased, to completely and accurately capture the information (Figure 8).

³ Analogue video captures less colour information than luminance information. The use of chroma subsampling (4:2:2) matches the digital encoding strategy to the analogue encoding strategy, and is acceptable. Lower resolution chroma subsampling, such as 4:2:0 and 4:1:1, are not acceptable for preservation.

⁴ It's important to remember the distinction being drawn between lossy compression, such as MPEG2, MPEG4, VC-1, Silverlight, IMX, etc., which are deemed unacceptable; and mathematically lossless compression such as JPEG2000, FFv1, and others. Lossless compression makes smaller files, but 100% of the data is recovered when decoded. The issue of added complexity, or loss of transparency, of these technologies is beyond the scope of the discussion of this paper.

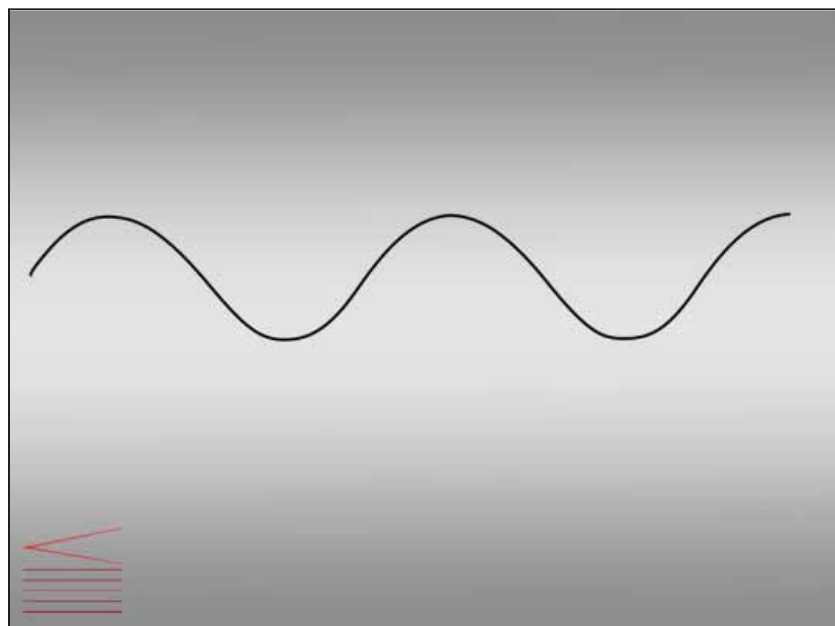


Figure 5. The simplest sound: sine wave.

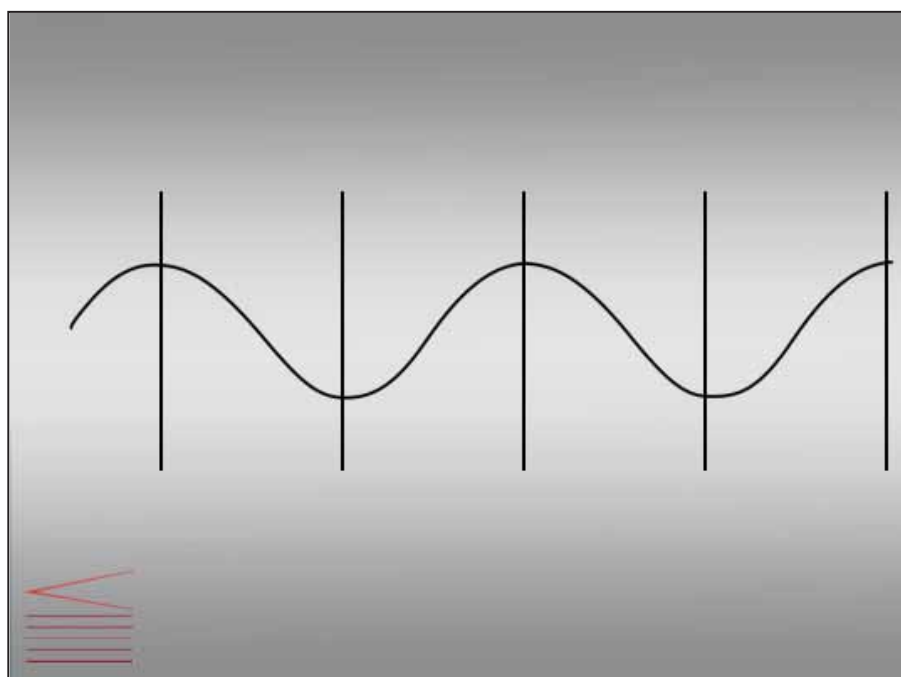


Figure 6. Sampling frequency determined by Nyquist formula.

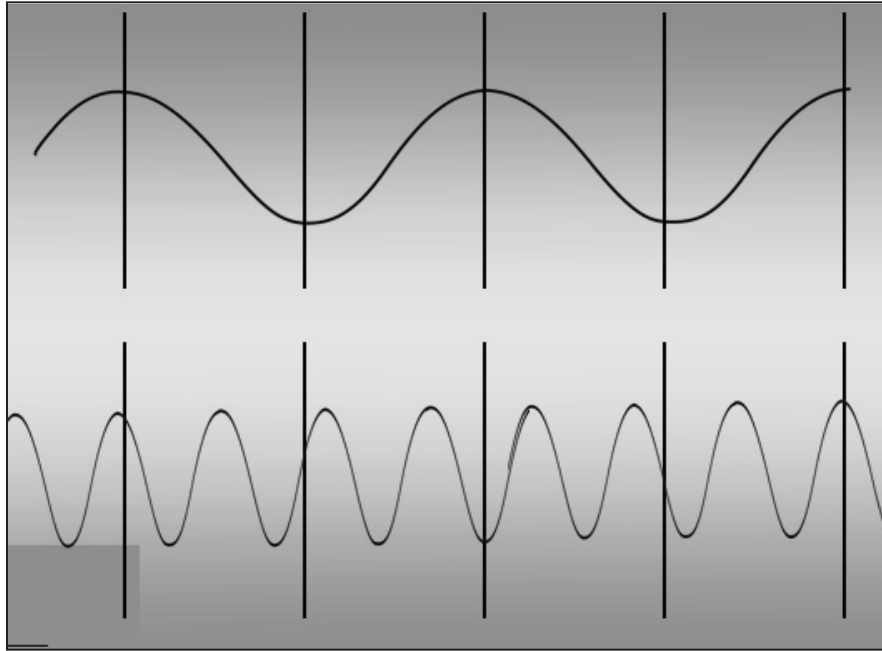


Figure 7. The sample rate at the top is adequate. On the bottom information between samples is not captured and lost forever.

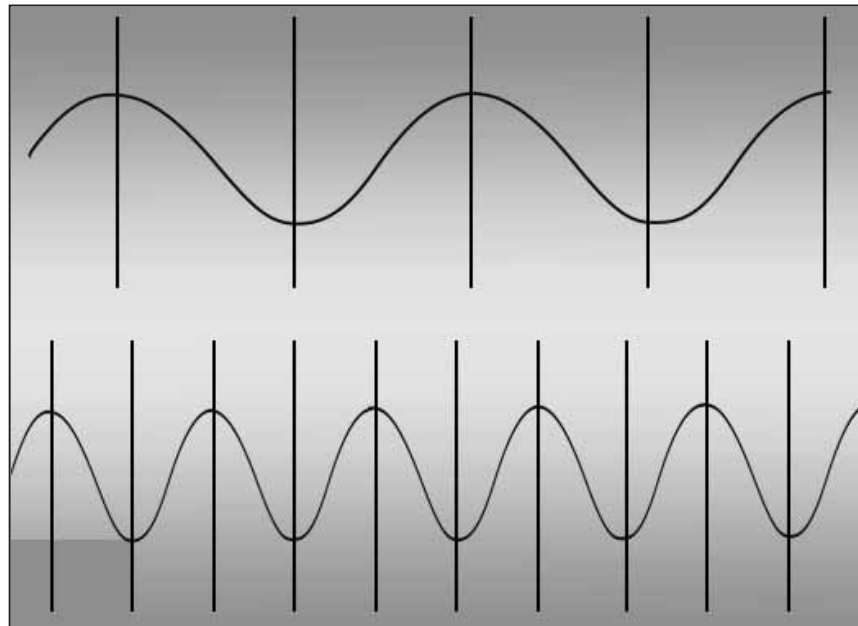


Figure 8. Higher rate (more bits) necessary to capture additional information in bottom example.

If we use the new higher sampling rate on the upper signal, no additional information is captured (Figure 9).

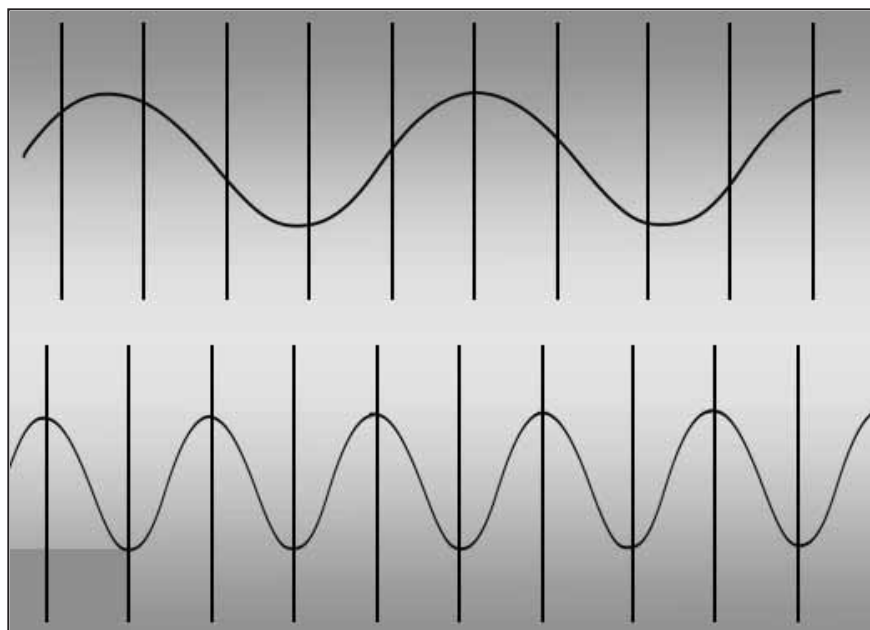


Figure 9. Higher rate used on bottom (higher quality) example doesn't capture additional information when used on top (lower quality) example.

Therefore, we can reduce the data rate, use fewer bits, by reducing the sampling for our upper example (Figure 10).⁵

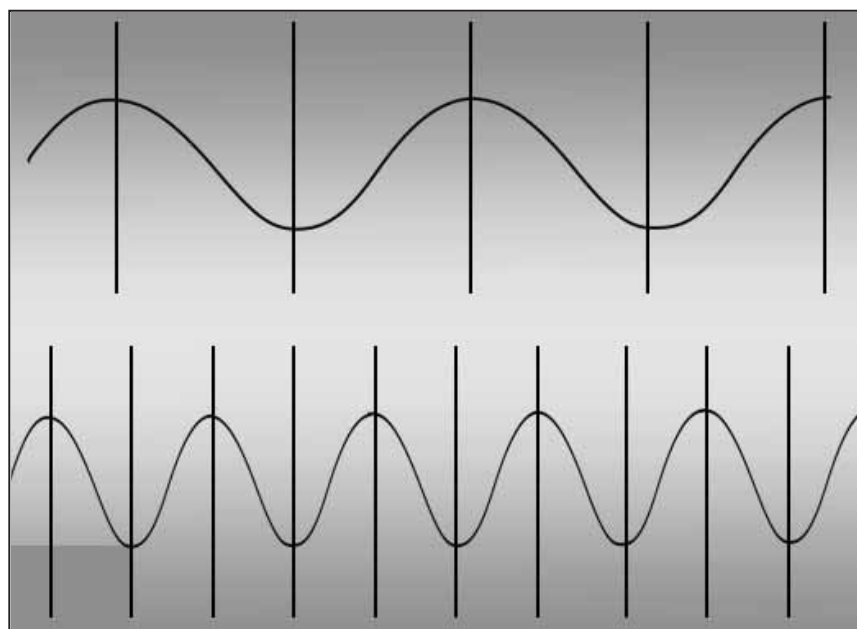


Figure 10. Lower bit depth can be used on lower quality, top, example.

⁵ This discussion is greatly simplified for clarity. The “real world” is more complicated. In practice, a sample rate higher than Nyquist is necessary. How much higher is not the point. The fundamental argument remains: a lower sampling rate is adequate when there is less high frequency information.

DATA RATE AND DIGITIZING VIDEO

Unlike analogue sound, which is a contiguous signal, analogue video is partially organized into discrete elements. Seconds are divided into frames, and the frames contain discrete horizontal lines. The lines, however, are a contiguous signal. When video is digitized, each frame is first sampled in what amounts to a TIFF image of each frame. There are 720 pixels across each of 486 lines.⁶

Let's say this is one line of video (Figure 11).

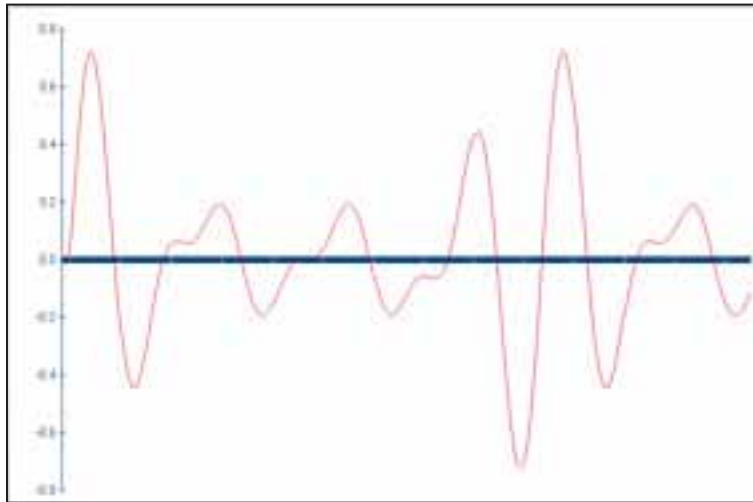


Figure 11. One line of video.

In essence, in uncompressed video, you do the same thing in audio. You sample each of the 486 lines 720 times (Figure 12).

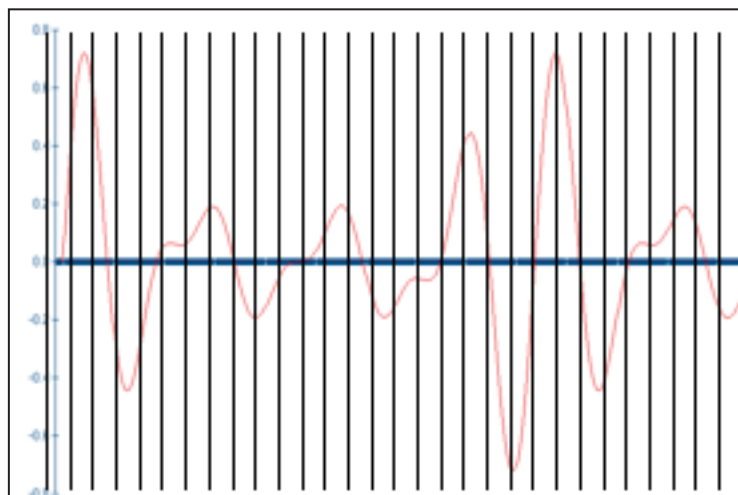


Figure 12. Line of video sampled at regular intervals.

⁶ The raster is 720 x 486 pixels in NTSC; and 720 x 576 in PAL.

This is the same process used to digitize still images, audio and video. At regular interval, in space and/or time, you capture a value.

- 600 pixels per inch (images)
- 96,000 samples per second (audio)
- 720 pixels per line (video)

There are 720 pixels across 486 lines of video, creating a grid of 720 x 486 pixels. When this is compressed, the first thing that happens is pixels are grouped into squares of 8x8 or 16x16 pixels.⁷ These are referred to as “tiles” or “macroblocks.”⁸ It is at this point, the very first stage of video compression, where compression becomes a bad thing: *compression fundamentally alters the organization of the original video* (Figure 13).

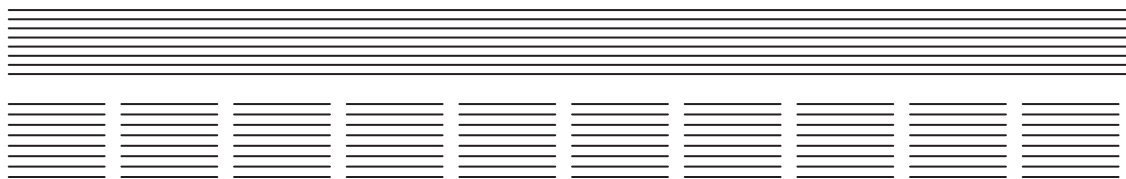


Figure 13. Lines are contiguous (top). When compressed organization is changed (bottom).

In an analogue video image there are horizontal divisions (the lines), but there are no vertical divisions. This is not a theoretical problem equivalent to the issue of all digitization of analogue signals—digitization entails taking a contiguous signal and sampling it in discrete elements that do not exist. In video it is as though you had taken a page from a book, which also has clear horizontal organization,⁹ cut between the lines of text, and also cut vertically up the page. No amount of wheat starch paste and long-fiber Japanese paper is going to reconstruct the fibers, which were cut creating the little squares.

If you sew, and work with fabric that has a pattern, you know the importance of matching the pieces of fabric at the seams. However well you match the fabric, you will still have a seam (Figure 14).

When you work with the fabric, iron it, make a pleat, assemble a piece, you will be aware of the seam; likewise with video. If you create these artificial divisions, you will encounter them in editing, dissolves, cross fades, colour correcting, etc.

I know what you’ve been thinking: ‘Houston, we have a problem.’ Everyone in the room who made it through 3rd grade math has done some simple division. *You don’t get an even number of blocks.* Eight divides into 720 evenly. But it does not divide evenly into 486.¹⁰

$$720 / 8 = 90$$

$$486 / 8 = 60.75$$

⁷ Again the technical discussion here is simplified for clarity. The 8x8 pixel block is a classic strategy that has been superseded by more complex algorithms. The fundamental argument remains: the picture is subdivided.

⁸ Video captures three values at each pixel, luminance (B&W) and two chrominance values. In this discussion we’ll describe how an 8x8 block of only one value (say only the luminance layer) is encoded. Technically that is a block. The combination of the 8x8 luminance layer and the two 8x8 chrominance layers is a *macroblock*.

⁹ In western scripts. The same argument applies for vertically oriented languages.

¹⁰ This is a problem with NTSC. PAL divides evenly: $576/8=72$.



Figure 14. “Can ya see it?” Seam line of very carefully matched patterned fabric.

But don’t you worry. This problem has been solved. We will just throw away 6 lines of video!

$$480 / 8 = 60$$

Wouldn’t it make all preservation simpler if we were allowed to do things like this? Just guillotine the brittle edges of books! Just brush away the pesky dust from pastels! Just low pass filter scratchy records! If preservation is about capturing as much detail *as possible*, that any bit of information not captured during digitization is forever lost, why do we allow this argument which deliberately and malice-aforethought discard over 1% of the information? Would you accept deleting 1 page out of every 100 in a book?

This image shows catastrophic macroblock decoding errors and it demonstrates the size of the macroblock units (Figure 15). During compression, the continuous image is broken into squares that have



Figure 15. Catastrophic macroblock decoding errors.

no relationship whatsoever to the original image.¹¹ These unrelated squares are encoded separately. Then they are reassembled, or concatenated, on playback to reconstruct the image.

Consider the line of video image again. Each line is sampled at 720 points and at each of these points each pixel is assigned a 10-bit value (Figure 16).

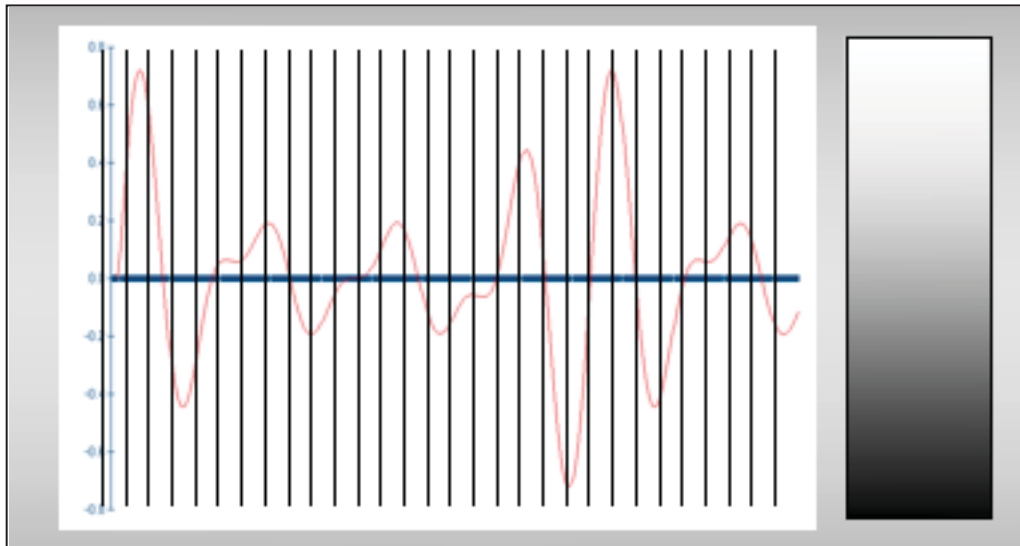


Figure 16. Line of video sampled at regular intervals. At each sample a 10-bit value is assigned.

Using compression, the signal is divided into segments (Figure 17).

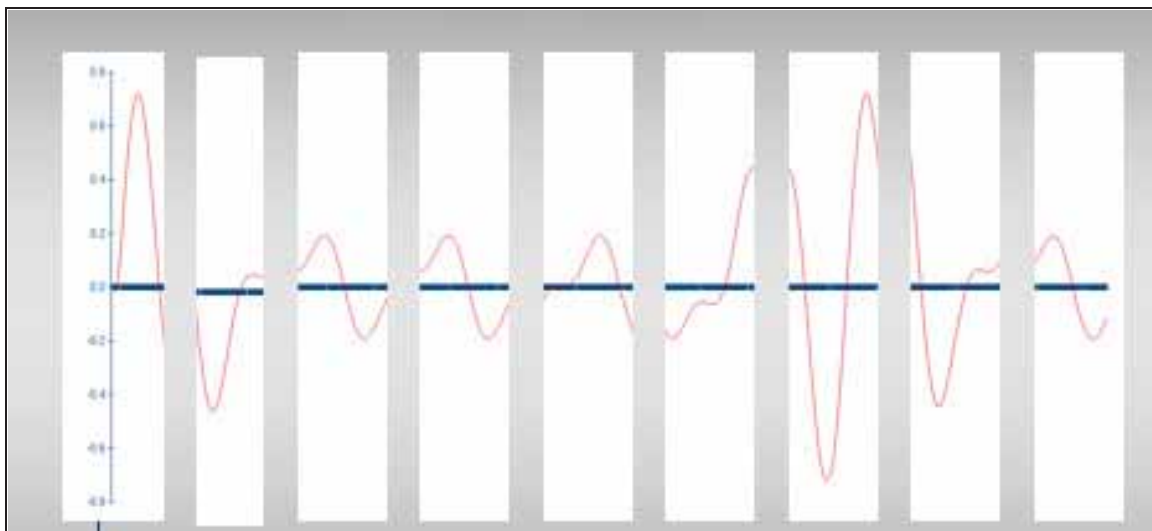


Figure 17. Effect on one line of re-organizing video for compression.

¹¹ As mentioned in footnote 5, this discussion is simplified for clarity. More recent codecs, such as MPEG4, contain strategies to address this problem using variable macroblock sizes. If an 8x8 block contains very different information, such as in this example the frame and the wall, or the matte and the picture, a different block size, such as 4x4, enables the subparts of the otherwise 8x8 block to be encoded separately, using a strategy better suited to the different parts. The image remains, however, artificially divided.

For each segment we write a formula that mathematically describes the waveform. You'll recall from high school algebra how you would take a formula, plug in a value for x then solve for y . You then take your sharpened #2 pencil and place a dot of the (x,y) pair on some graph paper. Then you'd plug in another value for x , get another y and repeat a few times. Finally you'd connect the dots, drawing a curve through the dots. Somehow it never looked quite right, your pencil was never quite sharp enough.

In encoding video we reverse the process. You start with the curve and derive a formula. This is a very simplified demonstration of how discrete cosine transforms (DCT) works. If you have a large amount of complex information, and video qualifies as a large amount of complex information, this is a more efficient way to represent the signal. However, there will always be some error. And the formulae here are complete nonsense. This example is just for illustration (Figure 18).

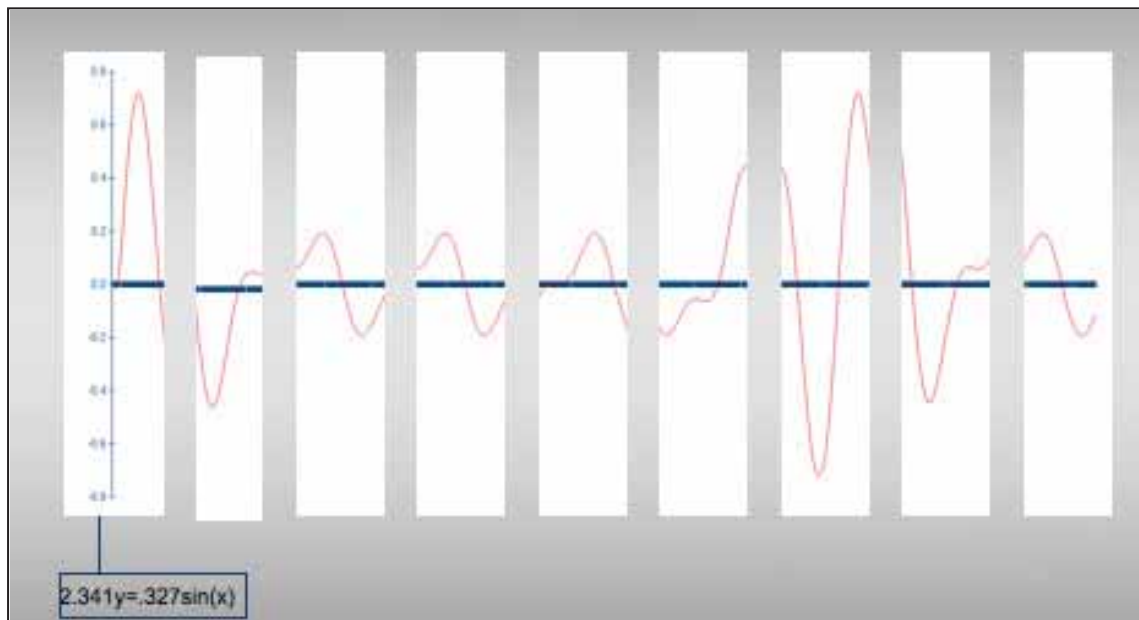


Figure 18. A formula is derived that describes a wave segment.

We do that for each segment of the video (Figure 19).

Just as your #2 pencil could never quite draw the perfect curve, we can never write a formula that matches the curve 100%. This creates two errors. The first is the difference between the analogue curve and the mathematical representation, called quantization error. There's an error in the quantity represented. The second error comes when the segments are stitched, or concatenated back together. The quantization error, however small, creates a discontinuity, an offset, when the wave segments are lined up next to each other.

The more resolution we have in our formula, the more closely it will approximate the waveform. In these examples, more places to the right of the decimal equals more resolution and require more bits.

$2.341y=.327\sin(x)$ has more information and uses more bits than $2.3y=.3\sin(x)$

When the data rate in video is reduced, the accuracy of the representation of the wave is reduced as well and the error at the seams is increased during concatenation. These are referred to as "concatenation errors at the macroblock boundaries."

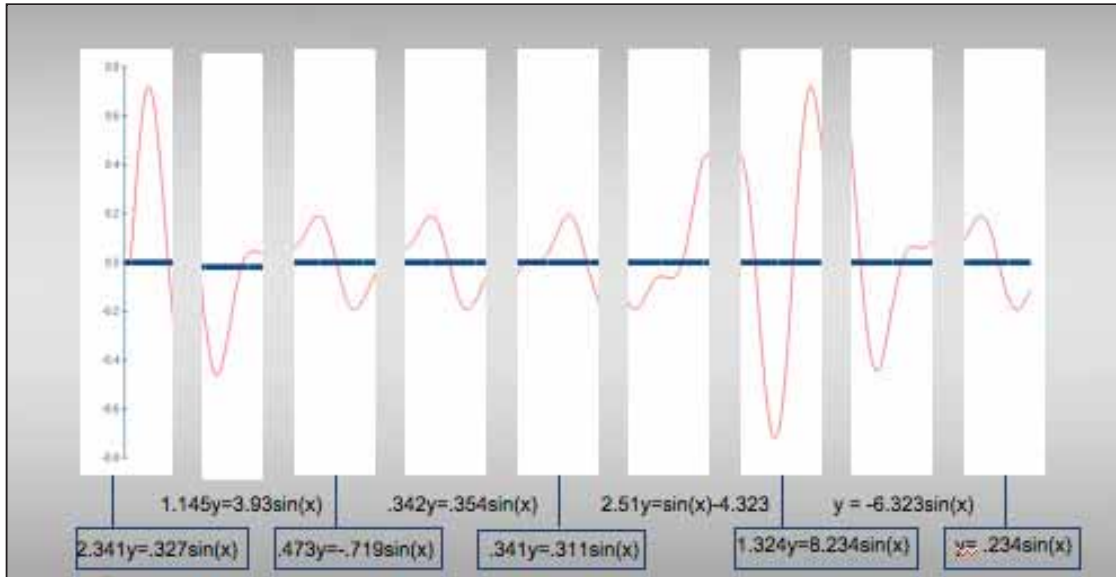


Figure 19. Formulae derived for each wave segment
(formulae in example are nonsense; just intended for demonstration).

Of course, it is more complicated than this. In the real world we are not trying to find a formula to describe a segment of just one line of video, but eight lines of video all at once (Figure 20).

This is happening both across and up and down the image. You have concatenation errors on all four sides of the macroblocks. *It is not a matter of whether you should use few bits for lower quality*

$$= X_k = \frac{1}{2}(x_0 + (-1)^k x_{N-1}) + \sum_{n=1}^{N-2} x_n \cos \left[\frac{\pi}{N-1} nk \right] \quad k = 0, \dots, N-1.$$

Figure 20. This is the discrete cosine transform formula;
much more complicated than the nonsense examples.

sources, but, by converting to macroblocks, it is not possible to turn the bit rate up high enough to overcome the change made by creating macroblocks. “Modern” or “intelligent” encoder designs are “smart enough” to adjust or adapt or interpolate for these concatenation errors. However, they cannot eliminate the fact that they have fundamentally changed the organization of the original signal by “smoothing” or “hiding” these errors. Like a finish carpenter installing trim in a living room, a tube of caulk hides a multitude of sins. All this leads to a softening of the image, the softness being an alteration of the original. Lower the data rate enough and you can see the artefacts very clearly. “Visually lossless” is a myth in preservation. Information has been discarded and forever lost. What may appear okay will impact all future uses of the video, from use in high quality productions to encoding the latest-most-used streaming format for the web.

EXAMPLE 1: <http://www.youtube.com/watch?v=NXF0ovsRcQg>

The fuzziness or softness of this image is the encoder working extremely hard to hide the macroblock tiles. Clearly the data rate is not high enough, but even when it is “high enough”, the structure of the original has been changed at the macroblock boundaries.

Transcoding from one compression to another can make this situation much worse. The cumulative error is extremely high and the image quality will suffer through multiple decode/encode cycles. MPEG2 uses 8x8 macroblocks; while MPEG4 uses variable block sizes that can be 4, 8, 12 or 16 pixels square. In transcoding, it is a certainty that the macroblocks will be subdivided differently for every frame and while this is good for encoder efficiency, it is very bad for transcoding between different compression schemata.¹²

This is an extreme example that uses the same JPEG encoding over and over, but it makes the case for cumulative encoding errors. Since it uses the same encoding algorithm, the macroblocking does not deteriorate.

EXAMPLE 2: <http://vimeo.com/3750507>

In an ideal world there would be commandments dictating the treatment of video for preservation. Perhaps something along the lines of:

- “Thou shalt not compress video”
- “If video is already compressed, you may leave it this way”
- “If you choose not to support this form of video compression, your only choice is to decompress and store in uncompressed” (with process history metadata telling it was previously compressed)

5. Conclusion

The history of the 20th century is unique in the amount of the human experience captured in time-based media, audio and video. The cultural record captured in video faces format obsolescence and rapid change. Playback equipment for legacy carriers and formats is rapidly disappearing and long out of production. This paper has argued that the race against the time when playback equipment will no longer be available should not lead to compromising fundamental principles of quality and accuracy of

¹² Indeed this is a large part of the problem in the YouTube example – it’s been transcoded a few times at low bit rates.

digitization. The technologies and formats exist to capture endangered media without significant loss. However many people advocate the use of lower resolution and compression in the name of smaller file sizes. The information lost in these decisions is forever lost and will negatively impact future use and access to this valuable material.

6. Recommended Target Formats for Digitizing Video

This is a summary of the recommendations for digital file formats for preserving video made in the white paper for the Library of Congress. The complete paper is available at:
http://dl.dropbox.com/u/11583358/IntrmMastVidFormatRecs_20111114.pdf

1. All analogue sources
 - 10 bit uncompressed, 720x486
2. Digital sources on tape: non-transcoded transfer possible
 - Keep native; may decode to uncompressed
3. Digital sources on tape: transcode necessary
 - 10 bit uncompressed, 720x486
4. Digital sources on other media
 - Evaluate: keep native or uncompressed
5. Optical discs
 - ISO Disc image

The Ibero-American Preservation Platform of Sound and Audiovisual Heritage

Pio Pellizzari,¹ Álvaro Hegewisch²

¹*Direttore Fonoteca Nazionale Svizzera, pellizzari@fonoteca.ch*

²*Director General de la Fonoteca Nacional de Mexico*

Abstract

The richness of sound and audiovisual collections in Latin America is noticeable, but so is also the risk of their imminent disappearance. The Fonoteca Nacional de Mexico, with the cooperation and support of IASA and FIAT promoted in 2010, a Latin American Meeting of Sound and Audiovisual Archives. The purpose was to establish a work group and to develop an integral project with a working agenda for rescuing sound and audiovisual archives at risk. A first experience has been made, in collaboration with the Fonoteca Nacional de Mexico and the CDI, analysing the archives of the indigenous radio stations in Mexico.

Author

Pio Pellizzari studied musicology, roman philology and French literature. He was a scientific collaborator for musicology at the libraries of the Universities of Lausanne and Fribourg elaborating musical inheritance and producing catalogues of musical works. He taught music at the secondary school, and was scientific librarian at the University of Fribourg, creating a sound archives for the University. In 2003/04 he has been invited professor at the University of Zürich. Since 1998, he is director of the Swiss National Sound Archives; board member of the National Museum of Switzerland; and since 2005 Vice-President of IASA and chair of the Training & Education Committee, which he founded.

Álvaro Hegewisch is a lawyer, cultural promotor and theater artist. He is currently Vice President of the International Association of Sound and Audiovisual Archives (IASA). He has been Chief Manager for special projects from the National Center for the Arts; General Director of Cultural Liasons; Vice Minister of Culture from México and General Director of the Fonoteca Nacional. He developed a National Audio Libraries Network in México, the IBERMEMORIA program for the preservation of sound and audiovisual collections in Iberoamerica, and the Training for Trainers program in collaboration with the Sound Archive of Switzerland and the Phonogramarchive of Austria and IASA.

1. Introduction

It is an undisputed fact that Latin America possesses an abundant audio and audiovisual heritage. But the richness of that heritage is matched by the dangers it faces, the risk of the imminent disappearance, it might in fact be lost forever. Many audiovisual collections are housed in highly precarious conditions, while the personnel responsible for them are often either badly trained or not trained at all. And the infrastructure necessary to ensure the preservation of this cultural legacy is insufficient, in some places non-existent.

There have been many efforts made up to the present day to counteract the possible disintegration of these collections, but these efforts were and remain often isolated and sporadic. In most cases they were only ever applied to one aspect of the conservation process. However, well-meaning they may have been, they have often remained without any degree of sustainability.

It is urgently necessary to define a common policy and strategy to develop common guidelines and a coordinated course of action in order to create the basic conditions needed to save this cultural heritage. This paper is intended to give a brief overview of a large-scale project initiated in Latin America. It will also offer an interim report on the activities undertaken up to now by taking a closer look both at smaller-scale projects that have already been completed and at the practical experience that has been gathered.

For several years, the *Fonoteca Nacional de Mexico* has been active in this regard. After Mexico created this institution for preserving audiovisual documents—the very first National sound archives in Latin America—it has in recent years developed into a real centre of competence for the region as a whole, taking on the necessary coordinating activities that come with such a function.

2. The First Steps

On the basis of its own experiences and in full knowledge of the current situation, the *Fonoteca Nacional de Mexico* initiated a Latin-American conference on sound and audiovisual archives in August 2010, in collaboration with IASA (International Association of Sound and Audiovisual Archives) and FIAT (Fédération Internationale des Archives de Télévision) and supported by both organisation.

The goal was to form a working group in which the different regions and countries are represented, and to fix an agenda in order to develop all-encompassing projects for endangered audiovisual collections. Six Latin-American countries participated: Mexico, Colombia, Costa Rica, Chile, Honduras, and Cuba and built up a strategic alliance for a collaboration. In order to achieve this goal, a preliminary, common plan of action was defined. This was to comprise the following points:

- Creating a questionnaire to make possible both an analysis of a collection and a diagnosis regarding its condition;
- Creating didactic guidelines for audiovisual archives;
- Creating a glossary in order to unify concepts and definitions;
- Making a didactic video to provide a preliminary aid for basic training;
- Drawing up a process model for developing a rescue plan;
- Realizing a pilot project in Mexico with the archives of indigenous radio stations.

During the first working meeting, conceptual references were standardised, and priorities and a working programme were defined. The first task on this programme was undertaken by the *Fonoteca Nacional de Mexico* together with IASA. They created a web platform, *The Ibero-American Preservation Platform of Sound and Audiovisual Heritage*, with the goal of providing tools to the different archives of the region to enable them to embark on preserving their cultural heritage in their own language and in accordance with their respective local circumstances.

The questionnaire is one of these tools. It is an aid to evaluation and description in the task of analysing an audiovisual collection. It was drawn up after the example of a similar questionnaire made for the ICRT (Istituto Cubano de Radio y Television) in Cuba, and on the basis of experiences made in the European TAPE project. Such analyses should give one a basis from which to establish the state of conservation of the individual sound and audiovisual documents, and to determine which archives are particularly endangered.

In a second step, initial, urgent measures can then be undertaken. As a preliminary measure for the conservation of sound collections in Costa Rica and Honduras, a list was drawn up with specific advice that took also into account the respective local conditions. Further to this, a diploma training programme for documentalists of audiovisual objects has been developed, and a first course for 25 specialists in this

field was held. Finally, a Spanish translation was made of both the TC-04, *The Guidelines on the Production and Preservation of Digital Audio Objects*, and of the IASA cataloguing rules for audiovisual documents. In the meantime, a first international seminar for sound archives was organised and carried out by the *Radio Nacional de Columbia*, featuring lectures and tutorials. On a political level, a matter of particular importance was the creation of the 'Iberoamerican Network Sound Heritage' at the Fourteenth Ibero-American Culture Forum. Furthermore, the *Fonoteca Nacional de Mexico* has built up a strategic collaboration with the following countries: Columbia, Costa Rica, Honduras, Cuba, Switzerland, Austria, France and Spain.

In order to consolidate its role as a competence centre and to be able to take on corresponding tasks, the *Fonoteca Nacional de Mexico*, together with the *Teaching & Education (T&E) Committee* of IASA, has developed a specific training and continuing education programme for future trainers. The goal of this programme is to gather together existing expertise in matters of handling and archiving sound recordings, of digitization and the archiving of audiovisual documents in digital form. Then in special courses, this expertise will be conveyed to the institution's own trainers (in other words, it offers training for trainers). This is especially important, since it means that when working with archives of indigenous peoples (such as 'Indio radio stations'), training can be offered in the local language and with local people who know the state of the conditions on the ground. Not least of all, this can also help to guarantee a certain degree of sustainability.

In the first programme, the chief technician of the *Fonoteca Nacional de Mexico* has undergone in-depth, specialized training in the Austrian Phonogrammarchiv in Vienna and in the Swiss National Sound Archives. The technicians working in these two institutions prepared a special work programme that dealt with all possible problem cases and how to solve them. The training was completed when the chief technician himself led a practical workshop at the annual IASA conference. This form of training and continuing education is already bearing its first fruits: two technicians working in regional radio archives in Mexico has been trained in the *Fonoteca Nacional de Mexico* and they are at present in Switzerland and in Austria in order to consolidate their knowledge. These are the first steps undertaken during the last years.

3. Further Steps Planned and Medium-Term Goals

The creation of a bibliography and documentation in Spanish is planned, as is the Spanish translation of a selection of specific articles, mostly published in English, on different aspects of audiovisual archiving. As far as possible, the basic training should be offered in the form of an online training programme, mainly for a first general training of archivists and documentalists. The mentioned Preservation Platform will also be made available for the exchange of experiences and is intended to be added to, gradually, with recommendations for the archiving and conservation of audiovisual documents, and it is to be supplemented by a list with guidelines for creating an emergency plan for endangered collections. A further plan is to promote a supra-regional, global inventory of the Ibero-American sound and audiovisual heritage, on the basis of standards recognised and supported by IASA and FIAT and coordinated by Mexico, through the *Fonoteca Nacional del CONACULTA*. The 'train the trainer' programme started successfully last year in the field of audio and digital technology for the preservation of audiovisual heritage. It is now to be expanded and consolidated. The goal is to build up a network of competency throughout the whole of Latin America. Finally, an "Iberomemoria Fund" is to be created, in

collaboration with the “General Secretary for Iberoamerica,” in order to offer financial support for endangered archives in developing and emerging countries.

4. A First Progress Report

In October 2010, the *Fonoteca Nacional de Mexico* and the *Comisión Nacional para el Desarrollo de los Pueblos Indígenas* (CDI) started an analysis programme together with those responsible for the sound archives of Mexico’s indigenous radio stations. For this purpose, the above mentioned questionnaire was used. At the same time this served as a test to establish the usefulness of the questionnaire itself. In parallel with this, a basic course was held in archiving audiovisual documents. Working with the questionnaire proved effective, but it also showed that the glossary with concepts and definitions is necessary to understand it properly. The combination of analysis work with basic training also proved very helpful. Early results included the identification of several endangered collections and the subsequent introduction of preliminary conservation measures. A special emergency plan was developed for the XEZON radio station, whose archives were in a very critical condition. Here, in brief, is a historical review of the state of things.

In November 1991, XEZON went on the air for the first time with a test broadcast. Right from the start, this radio station became the “voice of the *Sierra de Zongolica*” and became particularly important for the local indigenous population. In 1998, torrential rains flooded a large part of the city of Zongolica, with buildings up to 1.20 m under water. This included the radio station. Not only its infrastructure and its sound archive were damaged but also the whole radio building was ruined. The radio station thereupon moved to the *Centro para el Desarrollo Coordinador de los Pueblos Indigenas* in a higher part of the city of Zongolica. Today, it broadcasts every day from six o’ clock in the morning to six o’ clock in the evening. Its programmes are in Spanish and Nahuatl and cover the regions of Veracruz, Oaxaca, Puebla and Tlaxcala, which together have some 30 larger urban areas and about 690,000 inhabitants, of whom more than 226,000 are counted among the indigenous population.

The founding of XEZON Radio has served to strengthen the language Nahuatl in a sustainable manner and has led to the creation of important cultural organisations such as folk music groups and dance groups. It has also helped to prevent traditional indigenous healing practices from being forgotten. The audiences have been reminded again of their own traditional music, a music that had been driven out by the commercial music industry. The radio has given renewed life to the following:

- Music for traditional customs and ceremonies;
- *Nahua* dances and *Nahua* compositions;
- *Nahua* tales and stories;
- Folk music and traditional music;
- Traditional medicine;
- etc.

5. Analysis and Emergency Measures

The Radio’s extensive sound archives were dried as well as possible after the floods, taken to their new location and stored there. The analysis of the collection made in autumn 2010, however, brought the following situation to light. The sound documents—in this case, tapes—showed signs of decay and were being attacked by micro-organisms and fungus (see Figures 1-3). The glued joints on the tapes had

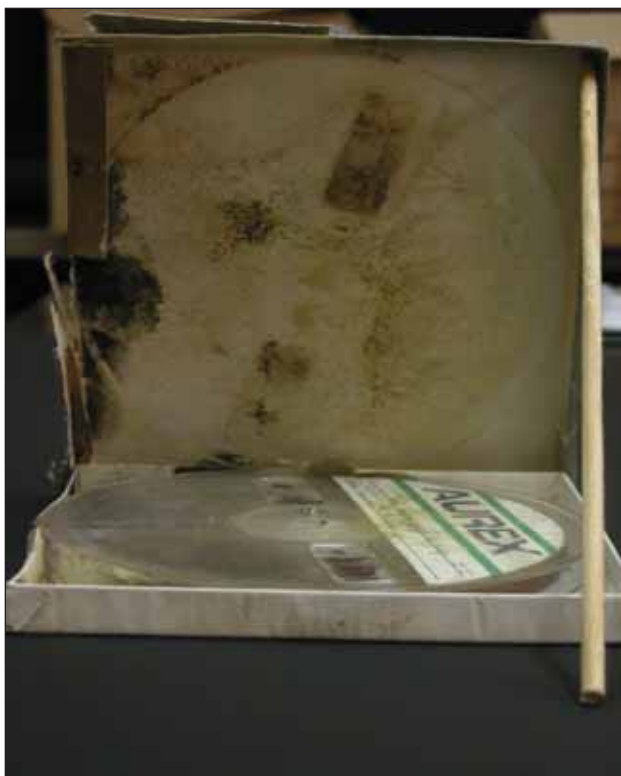


Figure 1.



Figure 2

crystallized and were beginning to come apart. The first attempts at cleaning proved to be quite difficult. Many tapes were damaged, with their reels deformed because of being stored both incorrectly and in unsatisfactory climatic conditions.

As a result of this analysis, the following steps for further work were determined:

- Transfer and admission of the collection to Fonoteca Nacional;
- Replacing all boxes and reels that had been infested by micro-organisms;
- Winding all tapes anew;

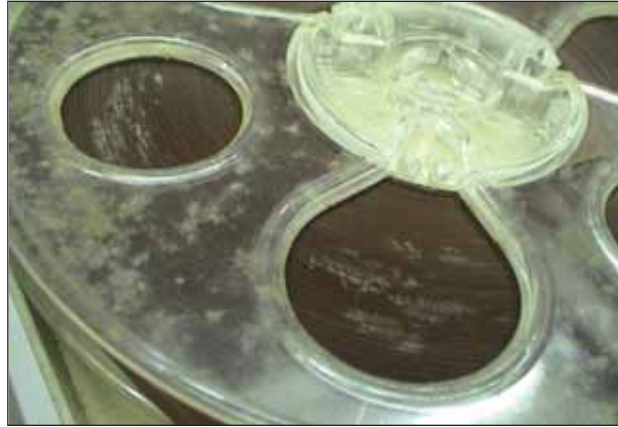


Figure 3.

- Stabilizing the tapes by placing them in a temporary storage space where the temperature and humidity were controlled;
- Giving an archive number to boxes and reels;
- Creating an inventory in order not to lose the original information.

When carrying out this work, it was discovered that the magnetic tapes were in an advanced state of oxidation, which meant that digitization had to be planned as quickly as possible. As a result, all the tapes were wound anew, and all old glued joints were taken apart, cleaned carefully and replaced by new ones. This work took place under the supervision of a specialist. Already here, the basic training that had been carried out at the beginning proved its worth.

In order to plan the digitization, a priority list was drawn up on the basis of the physical state of the tapes, the brand of tape, and in part, on the basis of their content. When digitizing, tests and quality control measures were carried out for each individual tape. The results, the technical data and supplementary remarks were recorded on a technical form filled out by the personnel responsible for transferring the audio document.

These salvaged documents are a part of the collective memory of the people of Zongolica, hitherto transmitted only orally, comprising their language, their music and dances, their customs and rites, and not least of all their knowledge—in this case, their knowledge of traditional medicine.

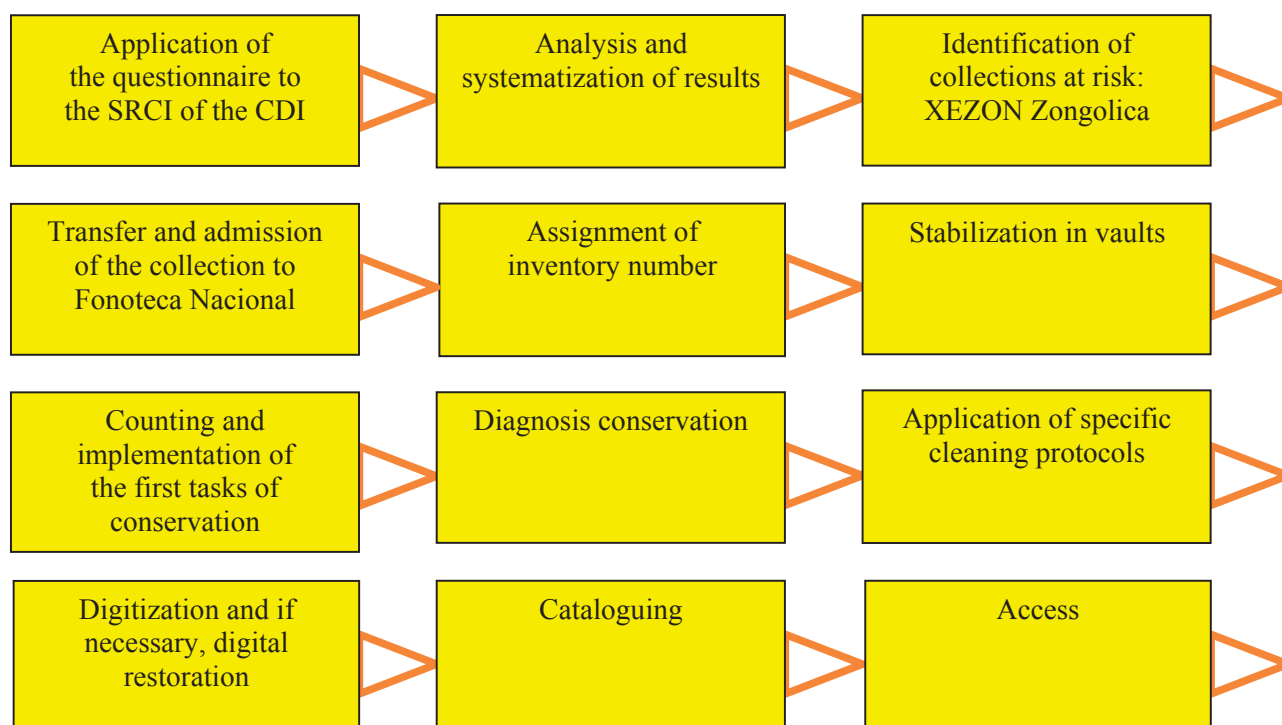
The conservation of over 400 magnetic tapes—their analysis, cleaning and copying into a digital format suited for long-term archiving—can serve as an example of how to develop a generic model for saving a sound collection. Above all, this project can also serve as a template for the indigenous communities of all regions of Latin America whose audiovisual heritage is in a similar state.

Figure 4 provides a summary of the workflow for the preservation of the Radio XEZON collection of Zongolica.

6. Future Prospects

Now our task is to forge further strategic alliances such as the one with the CDI in Mexico, which made these initial experiences possible. The collaboration with IASA and FIAT has to be promoted, enabling us to put those institutions' expertise to practical use by means of education and further training. Above all,

WORKFLOW FOR THE ATTENTION OF THE COLLECTION RADIO XEZON OF ZONGOLICA



however, we should win over further audiovisual archives in Latin America and convince them to join this platform in order to achieve the goals mentioned at the outset here:

- A comprehensive, global inventory of the audiovisual archives of Latin America;
- Expanding the training programme for future trainers; and
- Preserving the cultural heritage of Latin America according to recognised standards.

Figure 4.

This project enables the *Fonoteca Nacional de Mexico* to identify endangered sound and audiovisual archives in Latin America, to promote their conservation and thereby to help support the cultural identity of the local populations.

Trusting Data and Documents Online

Investigating and Analysing the Web-based Contents on Chinese Shanzhai Mobile Phones

Junbin Fang, Zoe Lin Jiang, Mengfei He, S.M. Yiu, Lucas C.K. Hui, K.P. Chow and Gang Zhou

Abstract

Chinese Shanzhai mobile phone has had a huge commercial market in China and overseas and was found to be involved in criminal cases. In this paper, a MTK-based Shanzhai phone with private web browser was investigated to extract user's web browsing data in the form of sites visited, received emails, attempted Internet searches and etc. Based on the findings, extracting Internet search conducted and web email received from the binary image was demonstrated. Besides, deleted browsing history can be recovered from snapshots in memory help reconstruct user's browsing activity and timeline analysis.

Authors

Dr. Junbin Fang is an associate professor in the Department of Optoelectronic Engineering and the Key Laboratory of Optoelectronic Information and Sensing Technologies of Guangdong Higher Education Institutes at Jinan University, Guangzhou, China. He is also working in the Department of Computer Science at The University of Hong Kong, Hong Kong, China. His research interests include mobile forensics, information security and quantum cryptography. junbinfang@gmail.com

Zoe Lin Jiang is a postdoctoral researcher in the School of Computer Science and Technology at Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China. Her research interests include digital forensics and applied cryptography.

Mengfei He is a master's student in the School of Computer Science and Technology at Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China. His research interests include digital forensic, information security.

Siu-Ming Yiu is currently an associate professor in the Department of Computer Science of the University of Hong Kong. His research interests include computer security, cryptography, digital forensics, and bioinformatics.

Lucas C.K. Hui is an associate professor in the Department of Computer Science, The University of Hong Kong, Hong Kong, China. His research interests include Information Security and Computer Forensics.

Kam-Pui Chow is an associate professor of Computer Science at the University of Hong Kong, Hong Kong, China. His research interests include information security, digital forensics, live system forensics and digital surveillance.

Gang Zhou is a director of computer application laboratory at the Wuhan Engineering Science & Technology Institute, Wuhan, China. His research interests include digital forensics, large scale network storage and embedded encryption system.

1. Introduction

The use of mobile phone has increased dramatically in the last decade. Globally, the number of mobile cellular subscriptions reached 5.3 billion in 2011, reported by the International Telecommunications Union (ITU). And vendors shipped 371.8 million units in Q1 2011, growing 19.8 percent year-over-year (IDC).¹ With the mobility and the portability, mobile phones have been part of people's daily life, which inevitably holds information of people's actions, whereabouts, and intentions. However, these advantages of mobile phone can be utilized by a criminal as a criminal tool anytime and anywhere, which leads to the necessity of mobile phone forensics. Mobile phone forensics is a branch of digital forensics which focuses on extracting information from a mobile phone as digital evidence in all kinds (criminal and civil) of court cases. With the fast evolution of mobile phone technologies, the amount and the types of data that can be found from a mobile phone are increasing. Traditionally, information that can be recovered from a phone includes phonebook, call logs, and short message service (SMS) messages,² even deleted items. Advanced smart phones also include wider varieties of data such as e-mails, media and web browsing data. The data can be stored in several storage media inside the phone, such as the SIM (Subscriber Identity Module) card, internal flash memory and external memory card.³ In terms of forensic, obtaining evidence from the internal flash memory is more challenging as SIM card and memory card can be taken down from mobile phone and therefore both of them can be investigated independently and deeply with external card reader.

Benefit from the "turn-key" development solution provided by MediaTek (MTK)⁴ and Spreadtrum,⁵ Chinese Shanzhai mobile phone (Shanzhai phone for short) has had a huge commercial market in China and overseas in recent years due to its high cost-performance ratios. Shanzhai phone is very cheap. For example, the price of a fake version of Apple's iPhone4S in the market is only \$130, and it can be down to \$60 for a cheaper (low-end) version. Unfortunately, with the worldwide spreading, more and more Shanzhai phones are found to be involved in criminal cases. However, little research findings on Shanzhai phone forensics has been published. One of the possible reasons may be that there is almost no existing official documents about the internal flash memory, the file systems and other related information for Shanzhai phones. The other reason may be that researchers did not expect the huge potential market of such low-price mobile phones so that little attention has been paid to it before.

In the paper, we provide the first step towards the forensics investigation on the web-based content on Shanzhai phone. Based on reverse engineering, we tried to provide important information of how a MTK-based Shanzhai phone manages and stores the web browsing data in its internal flash memory with its private web browser. This information provides insights on how to retrieve deleted web browsing history and helping the investigators rebuild the sequence of web addresses the user visited. The rest of the paper is organized as follows. Section 2 reviews the current work related to web forensics and Shanzhai phone forensics. Section 3 introduces the acquisition of Shanzhai phone's internal flash memory. Section 4

¹ Robin Wauters, "Worldwide Mobile Phone Market Grew 20% In Q1 2011, Fueled By Smartphone Boom," last modified April 28, 2011, <http://techcrunch.com/2011/04/28/worldwide-mobile-phone-market-grew-20-in-q1-fueled-by-smartphone-boom/>.

² Shafik G. Punja and Richard P. Mislan, "Mobile Device Analysis," *Small Scale Digital Device Forensics Journal* 2 (2008): 1–16.

³ S. Willassen, "Forensic Analysis of Mobile Phone Internal Memory," *Advances in Digital Forensics* 194(2005): 191–204.

⁴ MediaTek, "MediaTek," <http://www.mediatek.com/en/index.php>.

⁵ Spreadtrum, "Spreadtrum," <http://www.spreadtrum.com>.

details how to analyse and extract the web browsing data from the memory dump. Recovery of deleted browsing history and timeline analysis are described in Section 5. Section 6 concludes the paper.

2. Related Work

Traditionally, web forensics research is targeted to PC-based web browsers, such as Microsoft Internet Explorer, Mozilla Firefox, Google Chrom, Opera and Safari.

Jones and Rohyt described the IE and Firefox 2 Web browser forensics after simulating an actual crime in two different publications.⁶ Two free tools, the Pasco and Web Historian, were introduced for IE forensics with two commercial tools, the IE History and FTK tools. Forensics in Firefox 2 using a cache file and an analysis method using the cache file structure were also suggested in the publications. Pereira⁷ explained in detail the changes in the history system that occurred when Firefox 2 was updated to Firefox 3 and proposed a new method of searching deleted history information using unallocated fields. Oh et al.⁸ proposed an advanced evidence collection and analysis methodology for web forensics by performing integrated analysis across various browsers at the same time and using timeline analysis to detect the online movements of a suspect over time. A web forensics tool, named WEFA (Web Browser Forensic Analyzer), was also developed for the integrated analysis, which allows the investigator to examine the five leading web browsers (i.e., IE, Firefox, Chrome, Safari and Opera) existing in one system in parallel.

At present, there are a lot of tools for web forensics or forensics toolkits providing web browser investigation function, such as Netanalysis, Encase, FTK, etc. However, these methodologies and tools are originally designed for web browsers running on PC platform. When the scenario is moved to mobile phone platform, the investigation will become more challenging since the OS, the file system and the web browsers of the target are quite different with those on computer platform. Furthermore, the wide variety of mobile browsers also makes the forensic investigation more complicated.

Although there have been some mobile forensic toolkits which are dedicated to mobile OS, such as Android, iOS, Symbian and Windows Mobile, the toolkits cannot be used for Shanzhai phones since they are OS dependent. For Shanzhai phone forensics, Zhang⁹ discussed the recovery of MTK mobile phone flash file system, however, no detailed information is given. Fang et al.¹⁰ investigated how Shanzhai phone handles the addition and deletion of basic information in binary level as well as how to recover the historical data from memory image for timeline analysis. EDEC announced its Tarantula cell phone analysis system to target mobile phones using Chinese-manufactured chipsets and the system is further integrated into Logicube's CellXtract® cell phone forensic platform.¹¹ But both the systems didn't provide the function for web forensics.

⁶ Keith J. Jones and Blani Rohyt, "Web browser forensic," accessed January 19, 2012, <http://www.securityfocus.com/infocus/1827>.

⁷ Murilo Tito Pereira, "Forensic analysis of the Firefox3 internet history and recovery of deleted SQLite records," *Digital Investigation* 5 (2009): 93-103.

⁸ Junghoon Oh, Seungbong Lee, and Sangjin Lee, "Advanced evidence collection and analysis of web browser activity," *Digital Investigation* 8 (2011): S62-S70.

⁹ Zhi-wei Zhang, "The research of MTK mobile phones flash file system recovery," *Netinfo Security* 11 (2010): 34-36.

¹⁰ Junbin Fang et al., "Forensic Analysis of Pirated Chinese Shanzhai Mobile Phones," *Advances in Digital Forensics* VIII (2012): 117-130.

¹¹ Logicube, "Logicube Integrates EDEC's Tarantula To Target Mobile Devices Based on Chinese Chipsets," last modified March 04, 2012, <http://www.logicube.com/logicube-integrates-edecs-tarantula-to-target-mobile-devices-based-on-chinese-chipsets/>.

3. Internal Flash Memory Acquisition

Flash memory is currently the most dominant non-volatile solid-state storage technology for mobile phone. Similar to other brands of mobile phones, Shanzhai phones use flash memory as internal memory devices to store hard-coded system software, system files, user data, etc. Compared with reading data from the two other major storage units in mobile phones, i.e., SIM card and external memory card, retrieving data from internal flash memory is more complex and difficult as the memory chip cannot be read directly using external memory readers, except it is removed from the printed circuit board (PCB).

Nowadays, the methodologies for internal data collection can be classified into two approaches: physical and logical. The physical approach performs data extraction at a low level and allows obtaining the full memory image of contents of the entire phone memory, usually with the help of special hardware equipment. The logical approach uses communication protocols offered by the phone at a higher level, while the amount of acquired data is limited since the API provided by the phone were not developed for forensic purposes but to operate the phone as a modem. Another problem of logical approach is that contents deleted cannot be recovered in most cases.

As one of the physical approaches for internal memory acquisition, flasher tools may be the most convenient way to get a complete memory dump,¹² compared with the other two physical approaches—JTAG approach and physical extraction approach. Flasher tools are mainly used by mobile phone manufacturer and service providers to recover user data from dead or faulty mobile phones that otherwise will not provide access to data stored on their internal memory. They can also be used to update or replace software that is stored in the mobile phone. The forensic use of flasher tools is already being taught to future digital forensic examiners in Purdue's College of Technology in the United States of America.¹³ It is also being used by European investigators in mobile forensic cases.

In this paper, we go for the easier solution of using a flasher tool to obtain the memory dump instead of using JTAG or physical extraction since our focus is more on how the information is stored in the memory. In principle, the flasher tool connects with the UART interface (Rx and Tx pins) of Shanzhai phone's processor and run a serial communication protocol to communicate with the processor. When the power button of the Shanzhai phone is pressed, a boot loader inside the processor will be executed and can read/write from/to all registers and memory addresses, then the host software can successfully retrieve a full memory dump from the phone, as complete as JTAG approach does.

4. Analysing Web Browsing Data in Memory Image

Experiments were conducted on a typical model of Shanzhai phone, which is an imitated version of Apple's iPhone4. The model is equipped with a MediaTek MT6253 processor and a 16M byte NOR flash chip (Toshiba TC58FYM7T8C). After an image of the internal flash memory chip is dumped using flasher tool, the binary data will be analysed to extract related web information for forensic investigation.

MT6253 is Mediatek's first monolithic GSM/GPRS handset chip solution which offers highest level of integration with lowest power consumption and best-in-class features. Such that most of Shanzhai phone models were developed on this platform. The memory allocation scheme of the Shanzhai phone

¹² K. Jonkers, "The forensic use of mobile phone flasher boxes," *Digital Investigation* 6 (2010): 168-178.

¹³ Purdue University, "Expert: 'Flasher' technology digs deeper for digital evidence," accessed January 10, 2012, <http://phys.org/news95611284.html>.

under test followed the default configuration in MTK's turn-key solution. As shown in Figure 1, the Toshiba TC58FYM7T8C chip integrates 16 MB NOR flash memory and 4 MB RAM and the 16 MB flash memory was divided into two areas at offset 0xE0000 (corresponding to 14 MB). The first 14MB area was used as read-only memory (ROM) to store hard-coded system software, while the remaining 2MB area was used to store system files and user data. The 2MB area contained two drives in FAT12 format. One was for user data storage, which can be shown as a USB massive storage when the phone is connected to a computer. The other one was used as non-volatile random-access memory (NVRAM) for storing system files such as phonebook, SMS, call log, temporary files, etc. Noted that the NVRAM drive was invisible to normal user from computer or Shanzhai phone's UI.

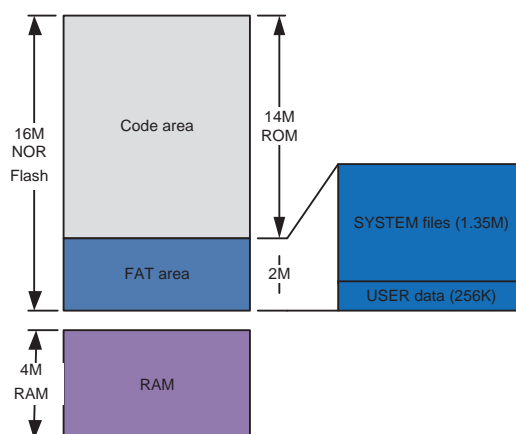


Figure 1. A typical memory allocation scheme of Shanzhai Phone.

4.1 Web browsing data extraction

Before going into the memory image to analyse the web-based information, the first step is to identify the type of the web browser used in the Shanzhai phone. If the web browser is one kind of the famous mobile browsers, the investigation would be easier as there may be a lot of previous research works and tools for the leading mobile browsers. Although it was reported at the end of 2011 that MediaTek and Opera Integrate Mini Browser on Feature Phones, the built-in browser of the Shanzhai phone is not Opera Mini but a private browser of Mediatek and there is not an existing tool can be applied to this browser. Since documents and references for the private browser are lack, reverse engineering is required to analyse the management of web browsing data, including browsing history, cache, bookmarks, cookies and etc.

In the experiments, after a series of common web browsing operations, a set of memory images were dumped from the Shanzhai phone and were further investigated, mainly the NVRAM area associated with the hard-coded area. Then, the organization of related files storing web browsing data in NVRAM was identified as follows:

- **Browsing history:** With the private browser, the browsing history is stored in two different files. The web address shown in address bar is classified into two categories, manual input and redirected links. The first kind of web address is stored under directory `"/bra"` with filename `"history.dat"`, while the last kind of web address is saved under the same directory with filename `"history2.dat"`. The format of a browsing history recorded in file `"history2.dat"` is as the example shown in Figure 2. The record is combined by two parts, header and body. The header field is 7

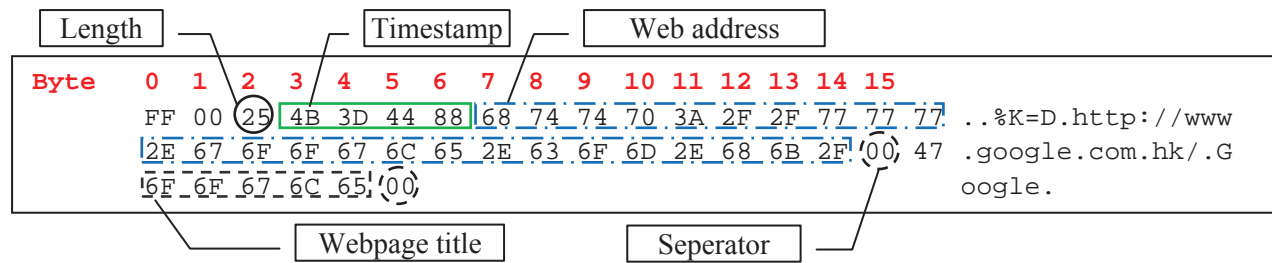


Figure 2. An example of a browsing history record in memory image.

bytes length with a “FF” start character and the third byte indicates the number of the remaining bytes of the record. The last 4 bytes in header is the Unix timestamp of accessing the website. The body field comprises the accessed web address and the title of the webpage visited, which are separated by a “00” byte. Note that the format for records in file “history.dat” is the same as that in Figure 2, except that there is not webpage title field.

- **Cache:** When user accesses a webpage, the content of the webpage is retrieved to local as cached Internet files. In the Shanzhai phone, cached Internet file is renamed and placed under the directory “./stk/cache”. In the directory, there is a file named “index.dat”, which is used to manage and index all the cached files stored in this directory. Every time when the Internet files are saved, the index.dat file will be updated. An example of the file allocation table (FAT) of the cache directory is shown in Figure 3.

Create time & date

Byte	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
49	4E	44	45	58	20	20	20	44	41	54	20	18	00	D8	04	INDEX	DAT
21	3C	00	00	00	00	35	05	21	3C	63	01	00	0C	00	00	!<....5.!<c.....	
53	33	36	46	34	43	32	31	42	49	4E	20	00	00	09	05	S36F4C21BIN	
21	3C	00	00	00	00	09	05	21	3C	6C	01	9A	00	00	00	!<.....!<l.....	
53	41	34	45	42	35	30	39	48	54	4D	20	00	00	0D	05	SA4EB509HTM	
21	3C	00	00	00	00	0D	05	21	3C	6D	01	0C	02	00	00	!<.....!<m.....	
53	35	36	39	38	41	39	31	47	49	46	20	00	00	34	05	S5698A91GIF ..4.	
21	3C	00	00	00	00	34	05	21	3C	75	01	D8	01	00	00	!<....4.!<u.....	

Figure 3. An example of cached Internet files record in memory image.

- **Cookies:** Cookies of browsing history are saved in the directory “./stk/cookie/”. The management of cookies is similar to that of cached Internet files. In the directory, a file name “index.dat” is used to manage and index all the files storing Cookies in that directory.
- **Bookmark:** Bookmarks of the private browser were stored in a file named “BKM.dat” under directory “./bra” in the NVRAM area.

4.2 Analysing Internet Search History and Web-Based Email

The above findings are related to the management and basic format of the web browsing data in the Shanzhai phone. From the perspective of forensics, critical evidence can be found in the suspect's web browsing data, including not only websites visited, but also Internet searches conducted and web-based email. For example, in the case of Neil Entwistle, he was convicted of murdering his wife and baby daughter after forensic investigators found a Google search for “how to kill with a knife” in his computer's web history.¹⁴

In this section, two experiments of analysing Internet search history and web-based email were demonstrated.

Internet search is helpful for people to get information effectively. At present, there are several popular search engines, such as Google, Yahoo, Baidu, Bing, etc. To investigate the Internet search conducted on the Shanzhai phone, we first extracted all the entries of browsing history using the pattern of “FF*****http”. Then a web address history with a general HTTP URL structure of Google search engine was found. As shown in Figure 4, the URL string located in memory image is “http://www.google.com.hk/search?hl=zhTW&newwindow=1&sky=ee&ie=Big5&q=hacking+hijack&btnG=%e6%90%9c%e5%b0%8b”. In this string, the search words are the value after the variable *q*, revealing that the suspect searched on Google using keywords “hacking” and “hijack”, which implies that the suspect has been interested in these techniques.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	0123456789ABCDEF
00ECF070	63	6F	6D	2E	68	6B	2F	00	47	6F	6F	67	6C	65	00	FF	g
00ECF080	00	91	4B	3D	44	C1	68	74	74	70	3A	2F	2F	77	77	77	..
00ECF090	2E	67	6F	6F	67	6C	65	2E	63	6F	6D	2E	68	6B	2F	73	.google.com.hk/s
00ECF0A0	65	61	72	63	68	3F	68	6C	3D	7A	68	2D	54	57	26	6E	earch?hl=zh-TW&n
00ECF0B0	65	77	77	69	6E	64	6F	77	3D	31	26	73	6B	79	3D	65	ewwindow=1&sky=e
00ECF0C0	65	26	69	65	3D	42	69	67	35	26	71	3D	68	61	63	6B	e&ie=Big5&q=hack
00ECF0D0	69	6E	67	2B	68	69	6A	61	63	6B	26	62	74	6E	47	3D	ing+hijack&btnG=
00ECF0E0	25	65	36	25	39	30	25	39	63	25	65	35	25	62	30	25	%e6%90%9c%e5%b0%
00ECF0F0	38	62	2B	00	68	61	63	6B	69	6E	67	20	68	69	6A	61	8b+.hacking hija
00ECF100	63	6B	20	2D	20	47	6F	6F	67	6C	65	20	E6	90	9C	E5	ck - Google
00ECF110	B0	8B	00	FE	06	BA	00	00	00	00	00	00	00	00	00	00

Figure 4. Extracting Internet search history in memory image.

Web-based email is a typical web application and all the main functions can be operated online using a browser. Different with other kinds of web browsing data, web-based email usually involve more personal information. And web-based email can be cached only when the content of the email has been shown in the browser. For example, reading an email brings the message up in the browser and causes it to be cached, while sending an email does not since the browser doesn't display the sent mail, except that the user read the mail in sent box. In Figure 5 through Figure 7, an experiment of extracting web-based email from the binary image is demonstrated. Noted that two email accounts in qq.com were used to send and receive email, respectively.

To investigate the web-based email which was read on the Shanzhai phone, we first extracted all the entries of browsing history using the pattern of “FF*****http”, same as did in investigating Internet search. Then the URL of reading email in mail.qq.com can be found. As shown in Figure 5, the URL is:

¹⁴ AFP, “Briton Googled ‘how to kill’ days before murders: court,” accessed February 14, 2010, http://afp.google.com/article/ALeqM5hNX7099fMvF7_qHCpy1Bz4VhtR6A.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	0123456789ABCDEF
00F3A710	FF	00	B5	4B	3D	44	92	68	74	74	70	3A	2F	2F	77	39	..K=D.http://w9
00F3A720	34	2E	6D	61	69	6C	2E	71	71	2E	63	6F	6D	2F	63	67	4.mail.qq.com/cg
00F3A730	69	2D	62	69	6E	2F	72	65	61	64	6D	61	69	6C	3F	68	i-bin/readmail?h
00F3A740	69	74	74	79	70	65	3D	30	26	73	69	64	3D	36	68	64	ittype=0&sid=6hd
00F3A750	6D	39	53	4D	32	4A	72	61	38	6D	51	42	51	57	38	6A	m9SM2Jra8mQBQW8j
00F3A760	54	35	76	4E	77	2C	35	2C	7A	54	4F	54	7A	4B	44	79	T5vNw,5,zT0TzKDy
00F3A770	32	26	66	6F	6C	64	65	72	69	64	3D	31	26	74	3D	72	2&folderid=1&t=r
00F3A780	65	61	64	6D	61	69	6C	26	73	3D	26	6D	61	69	6C	69	eadmail&s=&maili
00F3A790	64	3D	5A	43	33	30	32	31	2D	41	5F	34	36	73	6D	52	d=ZC3021-A_46smR
00F3A7A0	33	73	53	45	72	70	6A	39	76	43	30	39	63	64	32	37	3sSErpj9vC09cd27
00F3A7B0	26	6C	70	3D	30	26	6C	70	67	3D	26	74	6F	3D	00	51	&lp=0&lpq=&to=.Q
00F3A7C0	51	E9	82	AE	E7	AE	B1	00	FE	00	05	00	00	00	00	00	Q.....

Figure 5. The browsing history for the received web-email.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	0123456789ABCDEF
00F3C520	50	0A	15	88	08	35	A6	DC	4B	3E	96	11	53	32	66	38	P....5..K>..S2f8
00F3C530	34	64	35	34	2E	77	6D	6C	00	68	74	74	70	3A	2F	2F	4d54.wml1.http://
00F3C540	77	39	34	2E	6D	61	69	6C	2E	71	71	2E	63	6F	6D	2F	w94.mail.qq.com/
00F3C550	63	67	69	2D	62	69	6E	2F	72	65	61	64	6D	61	69	6C	cgi-bin/readmail
00F3C560	3F	68	69	74	74	79	70	65	3D	30	26	73	69	64	3D	36	?hitttype=0&sid=6
00F3C570	68	64	6D	39	53	4D	32	4A	72	61	38	6D	51	42	51	57	hdm9SM2Jra8mQBQW
00F3C580	38	6A	54	35	76	4E	77	2C	35	2C	7A	54	4F	54	7A	4B	8jT5vNw,5,zT0TzK
00F3C590	44	79	32	26	66	6F	6C	64	65	72	69	64	3D	31	26	74	Dy2&folderid=1&t
00F3C5A0	3D	72	65	61	64	6D	61	69	6C	26	73	3D	26	6D	61	69	=readmail&s=&mai
00F3C5B0	6C	69	64	3D	5A	43	33	30	32	31	2D	41	5F	34	36	73	lid=ZC3021-A_46s
00F3C5C0	6D	52	33	73	53	45	72	70	6A	39	76	43	30	39	63	64	mR3sSErpj9vC09cd
00F3C5D0	32	37	26	6C	70	3D	30	26	6C	70	67	3D	26	74	6F	3D	27&lp=0&lpq=&to=

Figure 6. The cached file for the received web-email.



Figure 7. Screenshot of web browsing cache file "S2f84d54.wml".

`http://w94.mail.qq.com/cgi-bin/readmail?hitttype=0&sid=6hdm9SM2Jra8mQBQW8jT5vNw,5,zTOTzKDy2&folderid=1&t=readmail&s=&mailid=ZC3021-A_46smR3sSErpj9vC09cd27&lp=0&lpg=&to=`

From this URL, two useful values can be extracted as:

- The unique id for the web session (sid): 6hdm9SM2Jra8mQBQW8jT5vNw,5,zTOTzKDy2
- The unique id of the mail which was read (mailed): ZC3021-A_46smR3sSErpj9vC09cd27

As the URL and the unique ids are known, the next step is to find the related cache file for this URL in the memory image. Taking the URL as a keyword to search in the index file of Cache or search directly in the image, a map between the cached file of the URL can be found. As shown in Figure 6, the cache file for the reading email is “S2f84d54.wml”.

After the file “S2f84d54.wml” was carved and reconstructed, it can be displayed in a wml parser. As shown in Figure 7, the wml file is parsed and contains the following information:

- Email service provider: QQ mailbox
- Email sender: Andy1
- Email receiver: testone
- Time: 10:28AM on July 21, 2012
- Subject: how to become a hacker

Content: Thinking Like a Hacker... Besides, accounts for this email can also be found in the raw format of the wml file:

- Email sending account: 1910159914@qq.com
- Email receiving account: 2576406269@qq.com.

Combining the findings in Internet search history and email, we can deduce that the user had put some effort in this kind of technique.

5. Recovery of Deleted Browsing History and Timeline Analysis

Most of web browsers provide function for clearing web browsing data, such as the cache, browsing history, cookies and etc. Users may delete the logs of web browsing to protect their privacy. Additionally, some web browsers can be configured to reset web browsing data periodically. In web forensics, if a suspect has performed such function on his web browser to destroy the trace of his activity, investigation will be more difficult.

For PC-based web browsers, recovery of deleted web browser information depends on the data clean-up mechanism of the browsers. For example, in firefox, log files are reinitialized and only the temporary files used by SQLite in unallocated space of disk are possible to be carved and recovered.

For Shanzhai phones, it was found that there are multiple copies of user data in the binary image.¹⁵ The copies are created at different time points due to the erasure and allocation mechanism of flash memory. Since flash memory can only be “erased” a block at a time and cannot offer arbitrary random-access rewrite or erase operations, we observed that when the data in Shanzhai phone’s web browser is updated, the new data will be stored in a newly erased block and the old data will be left untouched in the

¹⁵ Fang et al., “Forensic Analysis of Pirated Chinese Shanzhai Mobile Phones.”

old block. This operation leaves multiple copies of data items until the memory space is revoked and overwritten by some other data.

From the viewpoint of forensics, this characteristic can be utilized to recover more information, even when the web browsing data in Shanzhai phone was deleted automatically or intentionally.

This section details the results of recovering deleted browsing history and analysing user's web browsing activity. In the experiment, the following web browsing operations were performed on the Shanzhai phone in sequence:

- Step 1: User started the web browser and typed a test URL “http://i.cs.hku.hk/~jbfang” into the address bar of the browser.
- Step 2: User opened the test webpage, which contains 3 links to Google, Yahoo and Wikipedia separately.
- Step 3: User clicked the link of “Google” in the test webpage and the browser jumped to the homepage of Google's search engine (“http://www.google.com.hk”).
- Step 4: User searched “hacking hijack” in Google and waited for the results returned.
- Step 5: User deleted two entries of browsing history related to Google.
- Step 6: User typed URL “www.ask.com” into the address bar of the browser and accessed the website.
- Step 7: User searched “hacking hijack” in ask.com.

After the operations, a memory image was dumped from the Shanzhai phone and was investigated. Using pattern matching, we found 9 memory segments for browsing history in the binary image. One memory segment corresponds to a copy of browsing history after one operation and can be viewed as a snapshot. The snapshots of browsing history after every step are shown in Figure 8 through 14, noted as S1, S2, S3, S4, S5, S6, S7, correspondingly.

The snapshots in Figure 8 through Figure 11 show that every newly added record of browsing history is simply appended to the old records and causes to a new memory segment for browsing history. Snapshot in Figure 12 demonstrates the memory operation mechanism for deleting entries in browsing history. Deleting browsing history records didn't really erase the deleted entries in the file, but marked the deleted entries as invalid with a terminal symbol “FE”. In this case, the deleted records of browsing history still exist in the logical file as well as in the previous snapshots, thus it can be recovered from both the current memory segment and the previous ones. However, in the snapshots in Figure 13 and 14, the cases are different. The deleted entries were overwritten by the newly added record. Therefore, the deleted record of browsing history can only be recovered from the previous snapshots.

For this experiment, since the deleted browsing history can be recovered in the snapshots and the records have timestamps, the timeline of the user's web browsing activity can be rebuilt quickly using the timestamp information, i.e., the sequence for the snapshots is: S1-S2-S3-S4-S5-S6-S7. Nevertheless, if the phone was set to a wrong time purposely, timeline analysis may not be so straightforward since the phone was set to a wrong time purposely, timeline analysis may not be so straightforward since the timestamp in the records may not be able to reflect the real sequence of the web addresses accessed. In this case, the relative position of the records in snapshots could be taken into consideration to help deduce the actual sequence of user's web browsing activities. For example, even the timestamps were unbelievable, we can recognize that the user activity of accessing “www.google.com” is before visiting “www.ask.com” according to the mechanism of Shanzhai phone for adding/deleting records in browsing history.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	0123456789ABCDEF
00EDCE00	FF	00	20	4B	3D	3E	17	68	74	74	70	3A	2F	2F	69	2E	.. K=>.http://i.
00EDCE10	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	66	61	6E	cs.hku.hk/~jbfan
00EDCE20	67	00	00	FE	07	AA	00	00	00	00	00	00	00	00	00	00	g.....

Figure 8. S1 - Snapshot after step 1 (timestamp: 01 Jan 2010 00:13:11).

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	0123456789ABCDEF
00EDA200	FF	00	20	4B	3D	3E	17	68	74	74	70	3A	2F	2F	69	2E	.. K=>.http://i.
00EDA210	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	66	61	6E	cs.hku.hk/~jbfan
00EDA220	67	00	00	FF	00	31	4B	3D	3E	19	68	74	74	70	3A	2F	g....1K=>.http://
00EDA230	2F	69	2E	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	/i.cs.hku.hk/~jb
00EDA240	66	61	6E	67	3F	74	3D	33	37	37	35	30	00	46	6F	72	fang?t=37750.For
00EDA250	65	6E	73	69	63	73	00	FE	07	76	00	00	00	00	00	00	ensics...v.....

Figure 9. S2 - Snapshot after step 2 (timestamp: 01 Jan 2010 00:13:13).

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	0123456789ABCDEF
00ED6E00	FF	00	20	4B	3D	3E	17	68	74	74	70	3A	2F	2F	69	2E	.. K=>.http://i.
00ED6E10	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	66	61	6E	cs.hku.hk/~jbfan
00ED6E20	67	00	00	FF	00	31	4B	3D	3E	19	68	74	74	70	3A	2F	g....1K=>.http://
00ED6E30	2F	69	2E	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	/i.cs.hku.hk/~jb
00ED6E40	66	61	6E	67	3F	74	3D	33	37	37	35	30	00	46	6F	72	fang?t=37750.For
00ED6E50	65	6E	73	69	63	73	00	FF	00	25	4B	3D	3E	1F	68	74	ensics...%K=>.ht
00ED6E60	74	70	3A	2F	2F	77	77	77	2E	67	6F	6F	67	6C	65	2E	tp://www.google.
00ED6E70	63	6F	6D	2E	68	6B	2F	00	47	6F	6F	67	6C	65	00	FE	com.hk/.Google..
00ED6E80	07	4E	00	00	00	00	00	00	00	00	00	00	00	00	00	00	.N.....

Figure 10. S3 - Snapshot after step 3 (timestamp: 01 Jan 2010 00:13:19).

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	0123456789ABCDEF
00ECD000	FF	00	20	4B	3D	3E	17	68	74	74	70	3A	2F	2F	69	2E	.. K=>.http://i.
00ECD010	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	66	61	6E	cs.hku.hk/~jbfan
00ECD020	67	00	00	FF	00	31	4B	3D	3E	19	68	74	74	70	3A	2F	g....1K=>.http://
00ECD030	2F	69	2E	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	/i.cs.hku.hk/~jb
00ECD040	66	61	6E	67	3F	74	3D	33	37	37	35	30	00	46	6F	72	fang?t=37750.For
00ECD050	65	6E	73	69	63	73	00	FF	00	25	4B	3D	3E	1F	68	74	ensics...%K=>.ht
00ECD060	74	70	3A	2F	2F	77	77	77	2E	67	6F	6F	67	6C	65	2E	tp://www.google.
00ECD070	63	6F	6D	2E	68	6B	2F	00	47	6F	6F	67	6C	65	00	FF	com.hk/.Google..
00ECD080	00	91	4B	3D	3E	46	68	74	74	70	3A	2F	2F	77	77	77	..K=>Fhttp://www
00ECD090	2E	67	6F	6F	67	6C	65	2E	63	6F	6D	2E	68	6B	2F	73	.google.com.hk/s
00ECD0A0	65	61	72	63	68	3F	68	6C	3D	7A	68	2D	54	57	26	6E	earch?hl=zh-TW&n
00ECD0B0	65	77	77	69	6E	64	6F	77	3D	31	26	73	6B	79	3D	65	ewwindow=1&sky=e
00ECD0C0	65	26	69	65	3D	42	69	67	35	26	71	3D	68	61	63	6B	e&ie=Big5&q=hack
00ECD0D0	69	6E	67	2B	68	69	6A	61	63	6B	26	62	74	6E	47	3D	ing+hijack&btnG=
00ECD0E0	25	65	36	25	39	30	25	39	63	25	65	35	25	62	30	25	%e6%90%9c%e5%b0%
00ECD0F0	38	62	2B	00	68	61	63	6B	69	6E	67	20	68	69	6A	61	8b+.hacking hija
00ECD100	63	6B	20	2D	20	47	6F	6F	67	6C	65	20	E6	90	9C	E5	ck - Google
00ECD110	B0	8B	00	FE	06	BA	00	00	00	00	00	00	00	00	00	00

Figure 11. S4 - Snapshot after step 4 (timestamp: 01 Jan 2010 00:13:58).

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	0123456789ABCDEF
00F31400	FF	00	20	4B	3D	3E	17	68	74	74	70	3A	2F	2F	69	2E	.. K=>.http://i.
00F31410	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	66	61	6E	cs.hku.hk/~jbfan
00F31420	67	00	00	FF	00	31	4B	3D	3E	19	68	74	74	70	3A	2F	g....1K=>.http:/
00F31430	2F	69	2E	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	/i.cs.hku.hk/~jb
00F31440	66	61	6E	67	3F	74	3D	33	37	37	35	30	00	46	6F	72	fang?t=37750.For
00F31450	65	6E	73	69	63	73	00	FE	07	76	4B	3D	3E	1F	68	74	ensics...vK=>.ht
00F31460	74	70	3A	2F	2F	77	77	77	2E	67	6F	6F	67	6C	65	2E	tp://www.google.
00F31470	63	6F	6D	2E	68	6B	2F	00	47	6F	6F	67	6C	65	00	FE	com.hk/.Google..
00F31480	07	4E	4B	3D	3E	46	68	74	74	70	3A	2F	2F	77	77	77	.NK=>Fhttp://www
00F31490	2E	67	6F	6F	67	6C	65	2E	63	6F	6D	2E	68	6B	2F	73	.google.com.hk/s
00F314A0	65	61	72	63	68	3F	68	6C	3D	7A	68	2D	54	57	26	6E	earch?hl=zh-TW&n
00F314B0	65	67	77	69	6E	64	6F	77	3D	31	26	73	6B	79	3D	65	ewwindow=1&sky=e
00F314C0	65	26	69	65	3D	42	69	67	35	26	71	3D	68	61	63	6B	e&ie=Big5&q=hack
00F314D0	69	6E	67	2B	68	69	6A	61	63	6B	26	62	74	6E	47	3D	ing+hijack&btnG=
00F314E0	25	65	36	25	39	30	25	39	63	25	65	35	25	62	30	25	%e6%90%9c%e5%b0%
00F314F0	38	62	2B	00	68	61	63	6B	69	6E	67	20	68	69	6A	61	8b+.hacking hija
00F31500	63	6B	20	2D	20	47	6F	6F	67	6C	65	20	E6	90	9C	E5	ck - Google
00F31510	B0	8B	00	FE	06	BA	00	00	00	00	00	00	00	00	00	00

Figure 12. S5 - Snapshot after step 5.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	0123456789ABCDEF
00F29000	FF	00	20	4B	3D	3E	17	68	74	74	70	3A	2F	2F	69	2E	.. K=>.http://i.
00F29010	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	66	61	6E	cs.hku.hk/~jbfan
00F29020	67	00	00	FF	00	31	4B	3D	3E	19	68	74	74	70	3A	2F	g....1K=>.http:/
00F29030	2F	69	2E	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	/i.cs.hku.hk/~jb
00F29040	66	61	6E	67	3F	74	3D	33	37	37	35	30	00	46	6F	72	fang?t=37750.For
00F29050	65	6E	73	69	63	73	00	FF	00	19	4B	3D	43	62	68	74	ensics....K=Cbht
00F29060	74	70	3A	2F	2F	77	77	77	2E	61	73	6B	2E	63	6F	6D	tp://www.ask.com
00F29070	2F	00	00	FF	00	43	4B	3D	43	66	68	74	74	70	3A	2F	/....CK=Cfhttp:/
00F29080	2F	77	77	77	2E	61	73	6B	2E	63	6F	6D	3A	38	30	2F	/www.ask.com:80/
00F29090	3F	74	3D	30	34	32	37	39	00	41	73	6B	2E	63	6F	6D	?t=04279.Ask.com
00F290A0	20	2D	20	57	68	61	74	27	73	20	59	6F	75	72	20	51	- What's Your Q
00F290B0	75	65	73	74	69	6F	6E	3F	00	FE	07	14	6B	79	3D	65	uestion?....ky=e
00F290C0	65	26	69	65	3D	42	69	67	35	26	71	3D	68	61	63	6B	e&ie=Big5&q=hack
00F290D0	69	6E	67	2B	68	69	6A	61	63	6B	26	62	74	6E	47	3D	ing+hijack&btnG=
00F290E0	25	65	36	25	39	30	25	39	63	25	65	35	25	62	30	25	%e6%90%9c%e5%b0%
00F290F0	38	62	2B	00	68	61	63	6B	69	6E	67	20	68	69	6A	61	8b+.hacking hija
00F29100	63	6B	20	2D	20	47	6F	6F	67	6C	65	20	E6	90	9C	E5	ck - Google
00F29110	B0	8B	00	FE	06	BA	00	00	00	00	00	00	00	00	00	00

Figure 13. S6 - Snapshot after step 6 (timestamp: 01 Jan 2010 00:35:50).

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	0123456789ABCDEF
00F5B800	FF	00	20	4B	3D	3E	17	68	74	74	70	3A	2F	2F	69	2E	.. K=>.http://i.
00F5B810	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	66	61	6E	cs.hku.hk/~jbfan
00F5B820	67	00	00	FF	00	31	4B	3D	3E	19	68	74	74	70	3A	2F	g....1K=>.http:/
00F5B830	2F	69	2E	63	73	2E	68	6B	75	2E	68	6B	2F	7E	6A	62	/i.cs.hku.hk/~jb
00F5B840	66	61	6E	67	3F	74	3D	33	37	37	35	30	00	46	6F	72	fang?t=37750.For
00F5B850	65	6E	73	69	63	73	00	FF	00	19	4B	3D	43	62	68	74	ensics....K=Cbht
00F5B860	74	70	3A	2F	2F	77	77	77	2E	61	73	6B	2E	63	6F	6D	tp://www.ask.com
00F5B870	2F	00	00	FF	00	43	4B	3D	43	66	68	74	74	70	3A	2F	/....CK=Cfhttp:/
00F5B880	2F	77	77	77	2E	61	73	6B	2E	63	6F	6D	3A	38	30	2F	/www.ask.com:80/
00F5B890	3F	74	3D	30	34	32	37	39	00	41	73	6B	2E	63	6F	6D	?t=04279.Ask.com
00F5B8A0	20	2D	20	57	68	61	74	27	73	20	59	6F	75	72	20	51	- What's Your Q
00F5B8B0	75	65	73	74	69	6F	6E	3F	00	FF	00	60	4B	3D	44	62	uestion?...`K=Db
00F5B8C0	68	74	74	70	3A	2F	2F	77	77	77	2E	61	73	6B	2E	63	http://www.ask.c
00F5B8D0	6F	6D	3A	38	30	2F	77	65	62	3F	71	3D	68	61	63	6B	om:80/web?q=hack
00F5B8E0	69	6E	67	2B	68	69	6A	61	63	6B	26	71	73	72	63	3D	ing+hijack&qsrc=
00F5B8F0	30	26	6F	3D	30	26	6C	3D	64	69	72	00	41	73	6B	2E	O&o=0&l=dir.Ask.
00F5B900	63	6F	6D	20	2D	20	57	68	61	74	27	73	20	59	6F	75	com - What's You
00F5B910	72	20	51	75	65	73	74	69	6F	6E	3F	00	FE	06	B1	00	r Question?.....

Figure 14. S7 - Snapshot after step 7 (timestamp: 01 Jan 2010 00:40:02).

6. Conclusion

This paper presents a preliminary work on the investigation of the web-based contents on a MTK-based Shanzhai mobile phone by analysing the memory dump extracted from the phone using a flasher tool. The analysis reveals important details about how the web browsing data is managed and stored with the private web browser of the Shanzhai phone. Based on the analysis, Internet searches conducted and web-based email received can be extracted from the binary image. Besides, valuable historical data pertaining to multiple snapshots and deleted browsing history can be obtained from a memory dump as long as the associated memory locations have not been overwritten. This information can be used to help rebuild the time sequence of user's web browsing activities.

A Framework of Network Forensics and its Application of Locating Suspects in Wireless Crime Scene Investigation

Junwei Huang,¹ Yinjie Chen,¹ Zhen Ling,² Kyungseok Choo¹ and Xinwen Fu¹

¹University of Massachusetts Lowell, USA; ²Southeast University, China

Abstract

We propose to classify network forensic investigations into three categories based on when law enforcement officers conduct investigations in response to cyber crime incidents. We define proactive investigations as those occurring before cyber crime incidents; real time investigations as those occurring during cyber crime incidents, and retroactive investigation as those occurring after cyber crime incidents. We present a holistic study of the relationship between laws and network forensic investigations and believe that this framework provides a solid guide for digital forensic research. With the guidance of this network forensic framework, we propose HaLo, a hand-held device transferred from the Nokia n900 smartphone for the real-time localization of a suspect committing crimes in a wireless crime scene. We collect only wireless signal strength information, which requires low-level legal authorization, or none in the case of private investigations on campus. We found that digital accelerator on a smartphone and GPS are very often rough for measuring walking speed. We propose the space sampling theory for effective target signal strength sampling. We validate the localization accuracy via extensive experiments. A video of HaLo is at <http://youtu.be/S0vMe02-tZc>. In this demo, we placed a laptop that was sending out ICMP packets inside one classroom, used HaLo to sniff along the corridor and finally located the laptop.

Author

Junwei Huang received both Bachelor's and Master's degrees of Computer Science and Theory from Computer Science Department, Hunan University, Hunan, China. He is in the fourth year of applying his Ph.D. degree in UMass Lowell under the supervision of Prof. Xinwen Fu. His research includes Network Security and Network Forensics.

1. Introduction

Digital forensics is the science of collecting, preserving analysing and presenting evidence from digital devices (e.g., desktop computers, PDAs, PADs etc.) used and/or accessed for illegal purposes. The derived evidence needs to be sufficiently reliable and convincing to stand up in court. Digital Forensics is one of the fastest growing occupations to fight against computer crimes and a practical science for criminal investigations.¹

There are various classifications of digital forensics based on different criteria. One classification is hardware forensics² and software forensics.³ The former examines hardware code/architecture and the latter

¹ "Digital Forensics," Wikipedia, last modified 15 May 2012, http://en.wikipedia.org/wiki/Digital_forensics; Mark Pollitt, "A History of Digital Forensics," in *Advances in Digital Forensics VI*, ed. Kam-Pui and Sjujeet Shenoj (Boston: Springer, 2010), 3-15.

² Pavel Gershteyn, Mark Davis, and Sjujeet Shenoj, "Forensic Analysis of BIOS Chips," in *ibid.*, pp. 301-314; Pavel Gershteyn, Mark Davis and Sjujeet Shenoj, "Extracting Concealed Data from BIOS Chips," in *ibid.*, pp. 217-230; Pritheega Magalingam et al., "Digital Evidence Retrieval and Forensic Analysis on Gambling Machine," in *Digital Forensics and Cyber Crime*, ed. Sanjay Geol (Berling Heidelberg: Springer, 2010), 111-121; Paul K. Burke and

examines electronic document to identify document characteristics, such as authorship.⁴ In our paper, we classify digital forensics into computer forensics and network forensics. The former focuses on single alone devices while the latter deals with networks of devices and dynamic network traffic information. We focus on network forensics, which is still a frontier area of digital forensics and requires a lot of thinking.

In the past three decades, law enforcement specialists and academic researchers have invested a great deal of efforts into digital forensics to fight cyber crimes.⁵ They developed new areas of expertise and avenues of collecting and analysing evidences. The process of acquiring, examining, and applying digital evidences is crucial to the success of prosecuting a cyber criminal. However, digital forensics is a cross-disciplinary field and it requires knowledge of both computing and laws.⁶ Academic researchers often lack the required background in the relevant areas of laws.⁷ Because of this, their research results often fail to conform to legal regulations. They may be unfamiliar with the real-world problems faced by forensic investigators and the constraints involved in solving them. In reality, the incorrect use of new techniques may result in the suppression of gathered evidences in court. For example, using specialized technology to obtain information without warrants may violate the Fourth Amendment, and the evidence gathered may therefore suppressed in court.⁸

Since the first Digital Forensics Research Workshop (DFRWS) in 2001, numerous frameworks for digital forensics have been proposed to guide research and investigation.⁹ These frameworks are not uniform. However, there are certain commons to most frameworks, such as systematic evidence

Philip Craiger, "Xbox Forensics," *Journal of Digital Forensic Practice* 1, no. 4 (2007): 275-282; Brian D. Carrier and Joe Grand, "A Hardware-Based Memory Acquisition Procedure for Digital Investigations," *Digital Investigation* 1, no. 1 (2004): 50-60.

³ Andrew Gray, Philip Sallis, and Stephen MacDonnell, "Software forensics: Extending authorship analysis techniques to computer programs," in *Proceedings of the 3rd Biannual Conference of the International Association of Forensic Linguists (IAFL)* (1997): 1-8, accessed June 27, 2012, doi:10.1.1.110.7627; Juola Patrick, "Authorship Attribution for Electronic Documents," in *Advances in Digital Forensics II*, ed. Martin Olivier and Sujeet Sheno (Boston: Springer, 2006), 119-130; Olivier de Vel, Alison Anderson, Malcolm Corney, and George M Mohay, "Mining e-mail content for author identification forensics," *ACM SIGMOD Record* 30, no. 4 (2001): 55-64.

⁴ Juola Patrick, *Authorship Attribution (Foundations and Trends in Information Retrieval)* (Boston: Now Publishers, 2008).

⁵ Pollitt, "A History of Digital Forensics," pp. 3-15.

⁶ Gary Palmer and Mitre Corporation, "A Road Map for Digital Forensic Research," Report From the First Digital Forensic Research Workshop (DFRWS), Utica, New York, August 7-8, 2001; Ricci S.C. Jeong, "FORZA – Digital forensics investigation framework that incorporate legal issues," *Digital Investigation* 3, supp. (2006): 29-36; Ashley Brinson, Abigail Robinson, and Marcus Rogers, "A cyber forensics ontology: Creating a new approach to studying cyber forensics," *Digital Investigation* 3, supp. (2006): 37-43.

⁷ Robert J. Walls, Brian Neil Levine, Marc Liberatore, and Clay Shields, "Effective digital forensics research is investigator-centric," in *Proceedings of the 6th USENIX conference on Hot topics in security* (Berkeley: USENIX Association, 2011), 11.

⁸ Walls et al., "Effective digital forensics research"; *Kyllo v. United States*, 533 U.S. 27 (2001).

⁹ Palmer, "A Road Map for Digital Forensic Research"; Mark Pollitt, "Computer Forensics: an Approach to Evidence in Cyberspace," in *National Information Systems Security '95 (18th) Proceedings: Making Security Real* (Darby: DIANE Publishing, 1996), 487-492; Mark Reith, Clint Carr, and Gregg Gunsch, "An Examination of Digital Forensic Models," *International Journal of Digital Evidence* 1, no. 3 (2002), accessed June 28, 2012, <http://www.utica.edu/academic/institutes/ecii/ijde/articles.cfm?action=article&id=A04A40DC-A6F6-F2C1-98F94F16AF57232D>; Robert F. Erbacher, Kim Christensen, and Amanda Sundberg, "Visual Forensic Techniques and Processes," in *Proceedings of the 9th Annual NYS Cyber Security Conference Symposium on Information Assurance* (2006), 72-80; Karen Kent, Suzanne Chevalier, Tim Grance, and Hung Dang, "Guide to Integrating Forensic Techniques into Incident Response," *NIST Special Publication NIST-SP* (2006): 800-86.

collecting procedures.¹⁰ It is also agreed that different laws are constrained to different areas (e.g., military, private entities, law enforcement).¹¹ Nevertheless, most frameworks focus on technical details rather than detailed laws to guide research and investigation. In reality, due to the legal constraints, many available strategies are not practical for law enforcement. As a result, legal restrictions may preclude several criminal investigations.

In this paper, we integrate the framework of network forensics with actual laws in order to build a bridge between academic research and law investigation. To better assist law enforcement and make research practical, detailed laws are considered in our framework. From the view of law enforcement, we classify digital forensic investigations into three parts based on when law enforcement officers conduct investigations in response to crime incidents. We define proactive investigations¹² as those occurring before crime incidents; real time investigations as those occurring during crime incidents,¹³ and retroactive investigations as those occurring after crime incidents. This classification in terms of incident timing helps us understand related laws since laws are different if the investigation timing is different. It is derived from our careful study of traditional crime investigations, constitutional and statutory laws and due processes. Currently, most law enforcement investigations are proactive/retroactive investigations. Real time investigation is a critical issue for law enforcement.

In this paper, we first present a refined framework of network forensics with the Constitution and laws of the United States. Under the guidance of the framework, we developed a wireless network forensic tool HaLo (**H**and-held forensic **L**ocalization kit) for law enforcement in real time investigation. HaLo is transformed from a Nokia N900 smartphone and locates a suspect target in a building with received WiFi signal strength (RSS) while the suspect is committing a crime. We collect only wireless signal strength information, which requires low-level legal authorization, or none in the case of private investigations on campus. The basic idea of localization is to collect wireless signal strength samples while walking. The position where the maximum signal strength is measured will be a good estimate of the suspect device's location. The key challenge of accurate localization via the hand-held device is that the investigator has to control his or her walking speed and collects enough wireless signal strength samples. We find that digital accelerator on a smartphone gives a very rough estimation of walking speed. GPS is not appropriate for indoor use or for measuring low velocity such as walking speed. Thus, we propose an effective wireless sampling theory for HaLo in forensic localization in a wireless network

¹⁰ Jeong, "FORZA"; Pollitt, Mark, "Six blindmen from Indostan," (paper presented in the First Digital Forensic Research Workshop (DFRWS), Utica, New York, August 7-8, 2001); Nicole Lang Beebe and Jan Guynes Clark, "A hierarchical, objectives-based framework for the digital investigations process," *Digital Investigation* 3, no. 2 (2005): 147-167; Pollitt, "Computer Forensics"; Reith et al., "An Examination of Digital Forensic Models"; Erbacher et al., "Visual Forensic Techniques and Processes"; Kent et al., "Guide to Integrating Forensic Techniques into Incident Response."

¹¹ Sarah Mocas, "Building theoretical underpinnings for digital forensic research," *Digital Investigation* 1, no. 1 (2004): 61-68; Palmer, "A Road Map for Digital Forensic Research"; Jeong, "FORZA"; Pollitt, "Six blindmen from Indostan"; Brinson et al., "A cyber forensics ontology."

¹² Daniel Allen Ray, *Developing a Proactive Digital Forensics System* (Alabama: University of Alabama, 2007); Gary R. Gordon, Donald J. Rebovich, Kyung-Seok Choo, and Judith B. Gordon, "Identity Fraud Trends and Patterns: Building a Data-Based Foundation for Proactive Enforcement," Technical Report submitted to Bureau of Justice Assistance, Washington, D.C. 2007.

¹³ Swagatika Prusty, Brian Neil Levine, and Marc Liberatore, "Forensic Investigation of the OneSwarm Anonymous Filesharing System," in *CCS '11 Proceedings of the 18th ACM conference on Computer and communications security* (2011), 201-214; Marc Liberatore, Brian Neil Levine, and Clay Shields, "Strengthening forensic investigations of child pornography on P2P networks," in *Co-NEXT '10 Proceedings of the 6th International Conference* 19 (2010), 1-12.

crime scene investigation. We validate the localization accuracy via extensive experiments. Our research on effectively sampling RSS fills the missing theory of using hand-held devices for accurate localization. To date, no research has answered the question of how slow we should walk in order to collect enough RSS samples for accurate localization. This paper answers this very question.

The rest of this paper is structured as follows. Related work is introduced in Section 2. Section 3 details the refined framework of network forensics. In Section 4, we introduce HaLo, provide the localization algorithm and present the experimental results. We conclude the paper in Section 5.

2. Related Work

Due to space limitation, we only review existing work most related to our paper.

2.1 Digital Forensics

(Gray et al. 1997) applied authorship analysis techniques to computer program code in the area software forensics. They proposed several principal aspects of authorship analysis. (Juola 2006) made a contribution on software forensics by identifying the authorship of electronic documents rather than traditional paper documents. By mining properties and styles from electronic documents, people may identify the authorship characteristics of a document.

In hardware forensics, (Gershteyn et al. 2006) found BIOS can contain hidden information and introduced how to extract concealed information from BIOS. (Burke and Craiger 2007) found Xbox consoles can be modified to run malicious codes and developed tools to extract such information for forensic investigation. (Magalingam et al. 2010) retrieved information from non-volatile EPROM chip embedded in gaming machines for evidence recovery. (Carrierr and Grand 2004) proposed a hardware-based procedure to obtain information from volatile memory.

(Pollitt 1996, 2001) initialized an abstract framework for digital forensics and provided a historical overview of digital forensics.¹⁴ (Mocas 2004) identified three investigation entities: law enforcement, military and business enterprise. She built a common process for each entity. But she recognized that the participating events, constraints and outcomes could be different. (Jeong 2006) involved laws in digital forensic framework. However, he only included the abstract law notion in his framework rather than detailed laws. Later, (Brinson et al. 2006) proposed more detailed frameworks for digital forensics with law issues. But they did not address detailed laws for academic researchers and law enforcement investigators. (Beebe and Clark 2005) proposed an objectives-based framework for digital forensic processes. (Carrier and Spafford 2004) presented a simple framework for the digital investigation process that is based on the causes and effects of events, and later they used a mathematical model to present frameworks/classifications for digital forensic investigation.¹⁵ (Ren 2004)¹⁶ proposed a framework for a

¹⁴ Pollitt, "A History of Digital Forensics."

¹⁵ Brian D. Carrier and Eugene H. Spafford, "Categories of digital investigation analysis techniques based on the computer history model," *Digital Investigation* 3, supp. (2006): 121-130.

¹⁶ Wei Ren, "A Framework of Distributed Agent-based Network Forensics System" (paper presented in Digital Forensic Research Workshop 2004, Baltimore, Maryland, August 11-13, 2004).

distributed agent-based network forensics system in DSRWS 2004. Later on (Ren and Jin 2005)¹⁷ subsequently designed a distributed agent-based real time network intrusion forensics system. (Daniel 2007) devised a proactive forensic system that predicts attacks and changed its collection behavior before an attack takes place.

(Erbacher et al. 2011) described digital forensics from a forensic investigator's point of view. They indicated that without understanding the actual forensic context and constraints, academic research has little or no impact in reality. (Liberatore et al. 2010) also developed proactive/real time forensic tools over a public p2p network for law enforcement investigators to apply without legal constraints.¹⁸

2.2 Localization Algorithms on Smartphone

In our study, we aimed to locate an arbitrary WiFi including APs. (Zengbin et al. 2011)¹⁹ built a smartphone-based system for locating WiFi APs in real time. They implemented the system on Android phones. By rotating the smartphone several times in a place and analysing the signal strength, they were able to locate the direction of the target AP. The smartphone WiFi adapter is transferred into a directional receiver with the holding human body as a signal shield. (Sen et al. 2012)²⁰ modified the idea for indoor environment. They built a system SpinLoc relying on the signal strength of the direct signal path. They extracted the direct signal path from the power-delay profile of a link, physical layer information that is exported by the Intel 5300 card. They then repeated the same process and achieved the same goal with higher accuracy.

3. Framework of Network Forensics

We will present the refined framework of network forensics in this section. We first carefully compare traditional crime investigation and network forensic investigation. We then clarify certain law terminology and finally build up the framework of network forensics with laws.

3.1 Traditional Crime Investigation v.s. Network Forensic Investigation

We present three scenes in each traditional investigation. The first traditional crime investigation scene involves a police officer patrolling on the street and deterring (potential) criminals. We classify this process as a proactive investigation (i.e., occurs before a crime incident). Imagine the following scene. A robbery is happening on the street and a police officer sees the robbery, stops it and arrests the criminal. Here, crime is happening. Thus, we call it real time investigation. Now imaging a third scene. The robbery happened and the robber has fled. The police officer talks with the victim or other witnesses and

¹⁷ Wei Ren and Hai Jin, "Distributed Agent-Based Real Time Network Intrusion Forensics System Architecture Design," in *AINA '05 Proceedings of the 19th International Conference on Advanced Information Networking and Applications* 1 (2005), 177-182.

¹⁸ Prusty et al., "Forensic Investigation of the OneSwarm Anonymous Filesharing System"; Liberatore et al., "Strengthening forensic investigations of child pornography on P2P networks."

¹⁹ Zengbin Zhang, Xia Zhou, Weile Zhang, Yuanyang Zhang, and Gang Wang, "I Am the Antenna: Accurate Outdoor AP Location using Smartphones," in *MobiCom '11 Proceedings of the 17th annual international conference on Mobile computing and networking* (2011), 109-120.

²⁰ Souvik Sen, Romit Roy Choudhury, and Srihari Nelakuditi, "SpinLoc: Spin Once to Know Your Location," in *HotMobile '12 Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications* 12 (2012), 1-6.

conducts an investigation to determine what happened. They then eventually arrest the criminal. We call this process as a retroactive investigation.

Cyber crime investigation is very similar to traditional crime investigation. Consider the following three similar scenes. In the first scene, the police search a P2P network and try to identify the owner of illegal material. We call this a proactive investigation as it involves preparing for the detection of a crime incident. In the second scene, there is a hacker attacking a company's network. A police officer gets the report and monitors the activities on the Internet. The police then trace the activities back to the hacker, if possible, and eventually arrest the hacker. Because the crime is happening during the investigation, we call it a real time investigation. Normally, this type of investigation is used to monitor and preserve income/outcome traffic during the cyber crime and conduct the traceback process if possible. In the final scene, the police get a call after the hacking event. Law enforcement read the logs from the IDS and firewall, check the connection logs from local Internet Service Providers (ISPs) and then try to reconstruct the past session. They will eventually track it back to the hacker if possible and then arrest the hacker. Since the investigation is after the crime incident, we call it a retroactive investigation. The basic framework of network forensic investigation is shown in Figure 1.

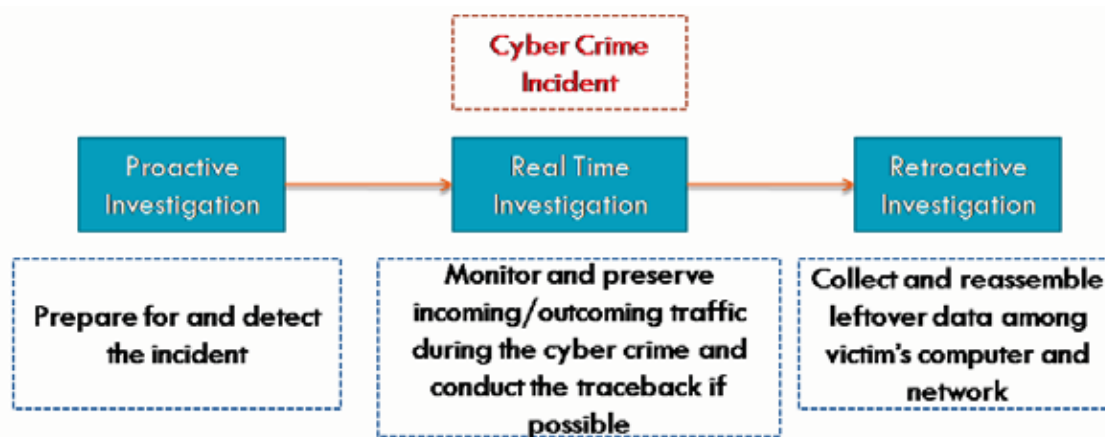


Figure 1: Basic Framework of Network Forensic Investigation.

Academic researchers normally develop tools for law enforcement in different investigations, but often ignore the legal constraints of such tools. Thus, it is difficult for law enforcement to use such kind of frameworks in actual investigations. Our framework, however, considers such legal constraints.

3.2 Terminology and Related Law Resources

Before addressing legal constraints in detail, we introduce relevant terminology and related legal resources in this section. Normally, there are two kinds of actions in cyber crime criminal investigations: investigations with warrants/court orders/subpoenas and investigations without warrants/ court orders/ subpoenas. They are governed by two primary law resources: the Fourth Amendment to the U.S. Constitution, and the statutory laws codified at 18 U.S.C. (United States Code) §§ 2510 to 2522, 18 U.S.C. §§ 2701 to 2712, and 18 U.S.C. §§ 3121 to 3127. Most cases involve either a constitutional issue under the Fourth Amendment or a statutory issue under the related law. In a few cases, they overlap.

3.2.1 Terminology

Subpoena: The process by which a court orders a witness to appear (and sometimes present information) in court and produce certain evidence. For example, law enforcement with a subpoena can require the witness ISP to produce connection logs to determine a particular subscriber's identity.

Court order: Official judge's statement compelling or permitting the exercise of certain steps by one or more parties to a case. For example, law enforcement can ask an ISP to install a packet-sniffer on its routers to collect all packets coming from a particular IP address to reconstruct an AIM session.

Search warrant: A written court order authorizing law enforcement to search a defined area and/or seize property specifically described in the warrant.

In general, the above processes are listed in order of degree of difficulty. For example, applying for a subpoena is much easier than applying for a search warrant. A mere suspicion is enough to apply for a subpoena, while "specific and articulable facts" are needed to apply for a court order and probable cause is necessary to apply for a search warrant.

3.2.2 Related Legal Resources

A. The Fourth Amendment to the U.S. Constitution

The Fourth Amendment is the main constitutional restriction to forensic investigation:

The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

The Fourth Amendment protects people's reasonable privacy by limiting government agents' authority to search and seize without a warrant. Government investigators cannot gather digital evidence and identify a suspect based on hunch; they must have probable cause.

B. Acts in United States Code (U.S.C.)

The following main restrictions from U.S.C. are also relevant.

1. Wiretap Act (Title III)

The Wiretap Act,²¹ 18 U.S.C. §§ 2510-2522, was first passed as Title III of the Omnibus Crime Control and Safe Streets Act of 1968 and is generally known as "Title III". It was originally designed for wire (see 18 U.S.C. § 2510(1)) and oral communications. The Electronic Communications Privacy Act of 1986 (ECPA)²² was enacted by the United States

²¹ "Wiretap Act," Wikipedia, last modified March 23, 2012, http://en.wikipedia.org/wiki/Wiretap_Act.

²² "Electronic Communications Privacy Act," Wikipedia, last modified May 24, 2012, <http://en.wikipedia.org/wiki/ECPA>.

Congress to extend government restrictions on wire taps from telephone calls to include transmissions of electronic data by computer.²³

The Wiretap Act is an important statutory privacy law. Roughly speaking, it prohibits unauthorized government access to private electronic communications (see 18 U.S.C. § 2510(12)) in real time.

2. Stored Communications Act

The Stored Communications Act (SCA),²⁴ 18 U.S.C. §§ 2701-2712, is a law that was enacted by the United States Congress in 1986. The SCA is a part of the ECPA. It protects the privacy rights of customers and subscribers of ISPs and regulates the government access to stored content and non-content records held by ISPs.

3. Pen Register Act

The Pen Register Act,²⁵ 18 U.S.C. §§ 3121-3127, is also known as the Pen Registers and Trap and Trace Devices statute (Pen/Trap statute). Generally speaking, a pen register device (see 18 U.S.C. § 3127(3)) records outgoing addressing information (such as a number dialed and receiver's email address); while a trap and trace device (see 18 U.S.C. § 3127(4)) records incoming addressing information (such as an incoming phone number and sender's email address).

In general, the Pen/Trap statute regulates the collection of addressing and other non-content information such as packet size for wire and electronic communications. Title III regulates the collection of the actual content of wire and electronic communications. Both of the two statutes above regulate the real-time forensics investigations while the SCA statute regulates the static forensics investigations (e.g., those involving email and account information). The relationship between network forensic investigations and laws is shown in Figure 2.

3.3 Reasonable Privacy

One critical concept in acquiring evidence is reasonable privacy. A person deserves reasonable privacy if 1) he/she actually expects privacy and 2) his/her subjective expectation of privacy is "one that society is prepared to recognize as 'reasonable'."²⁶ In this subsection, we discuss situations in which people have/do not have reasonable privacy.

²³ H. Marshall Jarrett and Michael W. Bailie, *Searching and Seizing Computers and Obtaining Electronic Evidence in Criminal Investigations* (Washington, DC: Office of Legal Education Executive Office, 2009), accessed June 28, 2012, <http://www.justice.gov/criminal/cybercrime/docs/ssmanual2009.pdf>.

²⁴ "Stored Communications Act," Wikipedia, last modified April 13, 2012, http://en.wikipedia.org/wiki/Stored_Communications_Act.

²⁵ "Pen Register Act," Wikipedia, last modified December 17, 2011, http://en.wikipedia.org/wiki/Pen_register#Pen_Register_Act.

²⁶ H. Marshall, *Searching*, 2009; EFF.org, "Reasonable Expectation of Privacy," accessed June 28, 2012, <https://ssd.eff.org/your-computer/govt/privacy>; *Katz v. United States*, 389 U.S. 347 (1967)

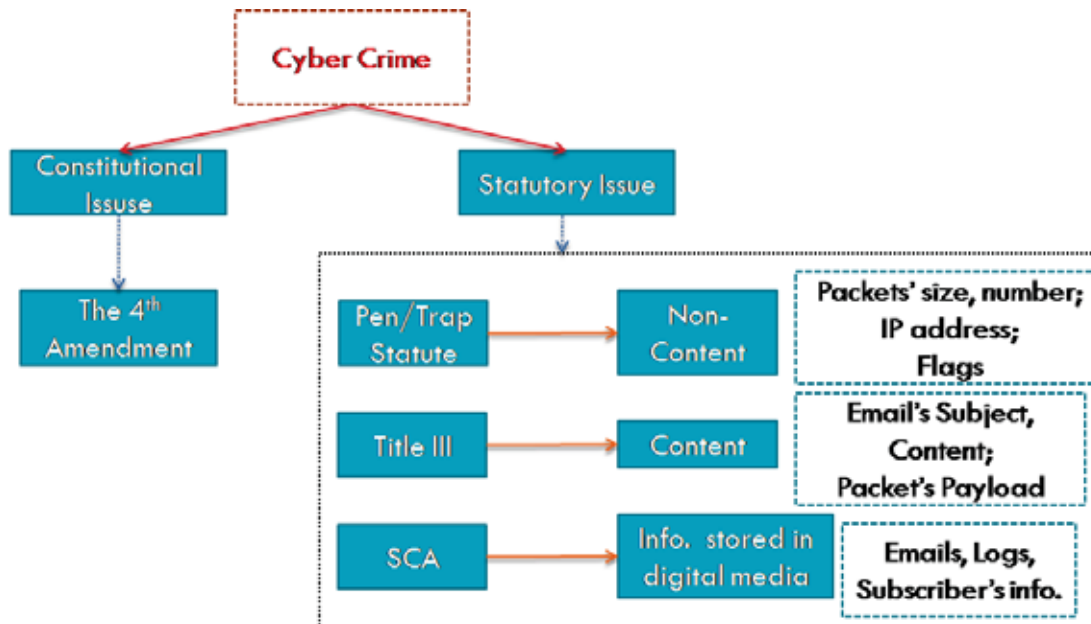


Figure 2: Relationship between Network Forensic Investigation and Laws.

3.3.1 When People have Reasonable Privacy

In 1967, the United States Supreme Court held that Katz, the defendant, had reasonable privacy when he entered a telephone booth, shut the door, and made a call. Thus, it was illegal for government agents to obtain the content of the phone call without a warrant, even though the recording device was attached outside the telephone booth, the communication was not interfered and the booth space is not physically intruded.²⁷ The Supreme Court holds that when the defendant shuts the door, his objective expectation is that nobody would hear his conversation and this action is recognized as reasonable by society. This idea is generally phrased as “the Fourth Amendment protects people, not places.”²⁸

A basic legal issue in digital forensics is whether an individual has a reasonable expectation of privacy of electronic information stored within computers (or electronic storage devices). The consensus is that electronic storage devices are analogous to closed containers and people do have a reasonable expectation of privacy. If a person enjoys a reasonable expectation of privacy of his/her electronic information, law enforcement officers ordinarily need a warrant to “search” and “seize”, or an exception to the warrant requirement before they can legally access the information stored inside. Therefore, when researchers invent a new technique, they need to determine whether this new technique violates a person’s expectation of reasonable privacy. If it does, they may need to re-design the technique in order to help law enforcement avoid search warrant requirements by searching for information not subject to privacy expectations.

²⁷ Katz v. United States, 389 U.S. 347 (1967)

²⁸ EFF.org, “Reasonable”, 2012

3.3.2 When People do not have Reasonable Privacy

Normally, individuals can have no reasonable expectation of privacy for information in public places. If a person knowingly exposes information to another person or in a public place, he/she has no reasonable expectation of privacy on that exposed information.²⁹ For example, two people are talking inside a house; they are talking so loudly that everyone walking outside the house can hear. Law enforcement on the street can record this conversation without a warrant, even though this conversation happens inside the house. In the *Katz* case,³⁰ although *Katz*'s conversation was not permitted to be recorded without a warrant, *Katz*'s appearance or actions (witnessed through the transparent glass) could be legally recorded. In other examples (e.g., bank accounts, subscriber information, the telephone numbers), there can be no expectation of privacy since the information is knowingly exposed to the service provider.³¹ However, that information is protected by statutory laws.

In digital forensics, if people share information and files with others, they normally lose the reasonable expectation of privacy. For example, a person has no privacy if he/she leaves a file on a public computer in a public library,³² or shares a folder with others.³³ Many cases have addressed sharing information and losing reasonable expected privacy, such as sharing information and files through P2P software³⁴ (including anonymous P2P software³⁵), leaving information on a public Internet³⁶ and so on.

Moreover, people may not retain their reasonable expectation of privacy if they relinquish control of the information and file to a third party.³⁷ For example, in digital forensics, a person may transmit information to third parties over the Internet or may leave information on a shared computer network. During the transmission, the government is not allowed to examine the content originally because it violates the both sender's and receiver's expected privacy.³⁸ The government needs a warrant to examine the information. However, the carrier of the information (e.g., the ISP) eliminates the privacy expectation (but that information is protected by statutory laws and the government still needs a warrant/court order/subpoena to obtain that information).³⁹ However, after the information is delivered, the sender no longer has a reasonable expectation of privacy (i.e., it "terminates upon delivery").⁴⁰

Another legal issue is that there is no agreement on whether a computer or other storage device should be classified as a single closed container or whether each individual file stored within a computer

²⁹ *United States v. Gorshkov*, 2001 WL 1024026, at *2 (W.D. Wash. May 23, 2001)

³⁰ *Katz v. United States*, 389 U.S. 347 (1967)

³¹ *Hoffa v. United States*, 385 U.S. 293, 302 (1966); *Smith v. Maryland*, 442 U.S. 735, 743-44 (1979); *Couch v. United States*, 409 U.S. 322, 335 (1973).

³² *Wilson v. Moreau*, 440 F. Supp. 2d 81, 104 (D.R.I. 2006); *United States v. Butler*, 151 F. Supp. 2d 82, 83-84 (D. Me. 2001).

³³ *United States v. King*, 509 F.3d 1338, 1341-42 (11th Cir. 2007); *United States v. Barrows*, 481 F.3d 1246, 1249 (10th Cir. 2007).

³⁴ *United States v. Stults*, 2007 WL 4284721, at *1 (D. Neb. Dec. 3, 2007).

³⁵ Swagatika, "Forensic," 2011.

³⁶ *United States v. Gines-Perez*, 214 F. Supp. 2d 205, 224-26 (D.P.R. 2002).

³⁷ *United States v. Horowitz*, 806 F.2d 1222 (4th Cir. 1986); *Guest v. Leis*, 255 F.3d 325, 333 (6th Cir. 2001); *United States v. Charbonneau*, 979 F. Supp. 1177, 1184 (S.D. Ohio 1997).

³⁸ *United States v. Villarreal*, 963 F.2d 770, 774 (5th Cir. 1992).

³⁹ *United States v. Young*, 350 F.3d 1302, 1308 (11th Cir. 2003).

⁴⁰ *United States v. King*, 55 F.3d 1193, 1196 (6th Cir. 1995); *Guest v. Leis*, 255 F.3d 325, 333 (6th Cir. 2001); *United States v. Meriwether*, 917 F.2d 955, 959 (6th Cir. 1990).

or storage device should be treated as a separate closed container.⁴¹ For example, if law enforcement wants to search a seized computer for child pornography, they may or may not use an exhaustive search tool to examine all files on this computer, while the owner of the computer may or may not have a reasonable expectation of privacy on some files which are not child pornography pictures. When researchers design such surveillance tools for law enforcement, they need to think about whether the tools violate the “reasonable expectation of privacy” of individuals.

3.4 Build up Framework of Network Forensics

In general, forensic investigators need a search warrant/court order/subpoena to pursue an investigation and gather the evidence legally. However, when the investigation does not violate a person’s reasonable privacy, does not break the law, or falls into an exception of law, then obtaining the evidence without a search warrant/court order/subpoena is not illegal, and the evidence will not be suppressed in court. Our previous work⁴² has presented this concept in detail, and thus, this will not be repeated in this paper.

In Figure 1, we classify the investigations into three categories based on when law enforcement officers conduct them. Proactive investigations occur before the crime incidents and are normally related to the Fourth Amendment. Law enforcement officers need to consider people’s reasonable expectation of privacy during investigations; otherwise, they may need a subpoena or court order. Real time investigations occur during the crime incidents and usually related to either statutory laws or constitutional laws. Title III and the Pen Register Act are used here in most cases. Normally, law enforcement needs a court order or search warrant to conduct such investigations. Retroactive investigations occur after crime incidents and are related to either statutory laws or constitutional laws, but the SCA is used here in most cases. In reality, law enforcement needs subpoena, court order, or search warrant, or all three to conduct investigations. The refined framework is shown in Figure 3.

Currently, law enforcement focuses on retroactive investigations for cyber crime because of legal restriction. Unlike the military or private entities, law enforcement cannot directly monitor the Internet because of privacy issues. Our research focuses on the development of forensic tools for law enforcement to conduct real time investigations. The best tools for law enforcement are those without any legal restrictions. However, in most cases, it is very hard to find such tools. In reality, network forensics investigations are a systematic process. In some cases, law enforcement may already have low-level authorization and they can use corresponding tools to conduct real time investigation and then obtain a high-level authorization to do in-depth investigation.

4. HaLo - Forensic Localization Tool

We studied a generic cyber crime scene: A suspect Bob is stealing his neighbor’s (Alice) WiFi and doing illegal activities such as downloading child pornography movies. Law enforcement traces the activity

⁴¹ *Guest v. Leis*, 255 F.3d 325, 333 (6th Cir. 2001); *United States v. Runyan*, 275 F.3d 449, 464-65 (5th Cir. 2001); *United States v. Beusch*, 596 F.2d 871, 876-77 (9th Cir. 1979); *United States v. Walser*, 275 F.3d 981, 986 (10th Cir. 2001).

⁴² Huang, Junwei, Zhen Ling, Tao Xiang, Jie Wang, and Xinwen Fu, “When Digital Forensic Research Meets Laws,” (accepted by the First International Workshop on Network Forensics, Security and Privacy (NFSP 2012), Macau, China, June 18-21, 2012).

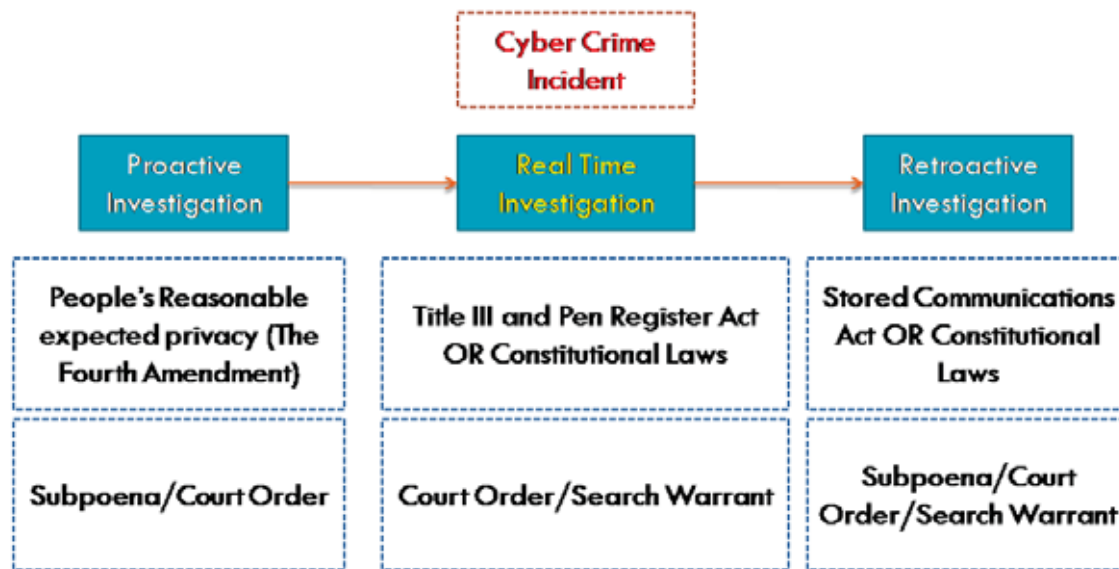


Figure 3: Framework of Network Forensics with Laws.

backs to Alice's router and obtains authorization to monitor the activities of Alice's router. However, since there is no information on Bob, law enforcement is unable to lock the suspect Bob. Law enforcement cannot break into Alice's neighbors' houses since they do not have search warrants for Alice's neighbors at this moment. Therefore, our aim was to design a tool for law enforcement to locate the suspect Bob. This scene is illustrated in Figure 4.

Since law enforcement has authorization to monitor Alice's router, law enforcement knows the suspect's (Bob's) MAC address. We designed a localization algorithm to locate Bob's physical location,

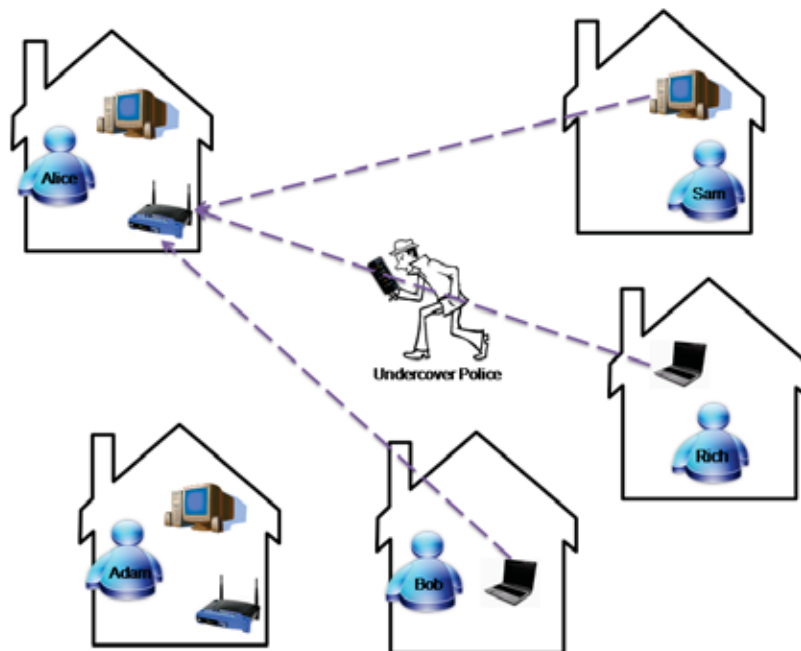


Figure 4: Cyber Crime Scene.

which requires Bob's signal strength and we used a Nokia N900 smartphone to detect the signal strength. The law enforcement agent walks next to each house or along a sideways corridor and collects the target's RSS. With RSS, the agent is able to locate the suspect Bob. Therefore law enforcement can then obtain a search warrant for Bob and later search his computer.

Figure 5 is the GUI of our forensic tool. By loading the bleeding-edge wl1251 driver for Maemo Fremantle,⁴³ the N900 device can work in monitor mode and is able to monitor any MAC address on any

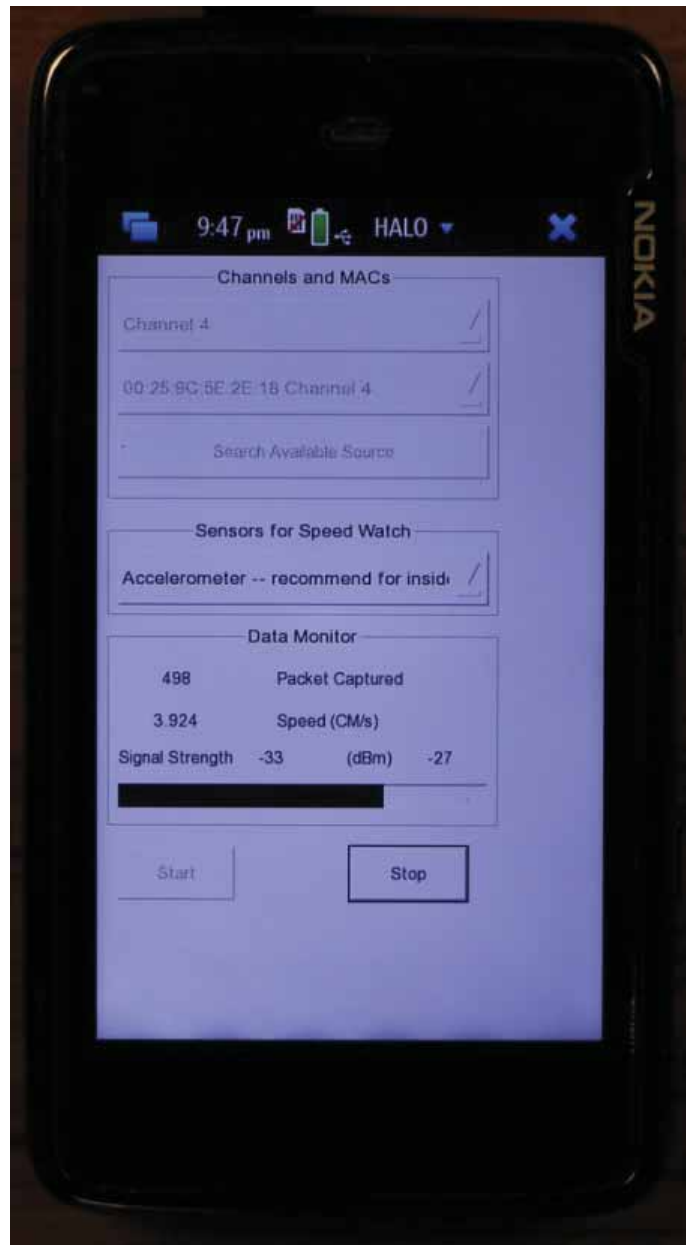


Figure 5: GUI of HaLo.

⁴³ David, "bleeding-edge wl1251 driver for Maemo Fremantle," *David's IT blog*, accessed June 28, 2012, <http://david.gnedt.eu/blog/wl1251/>.

channel. This tool is implemented with the libpcap library. Therefore it is able to capture packets from the target. There is an indicator at the bottom of this tool that indicates the maximum signal strength detected and the signal strength of current captured packet. We programmed this software using the Qt Creator. Thus, law enforcement can secretly monitor all connections with Alice's router. They must have a search warrant for Alice.

In case law the enforcement agent walks too fast and misses packets from the target, we implement two methods to estimate the device's moving speed. The agent can switch to GPS (outdoors) or Accelerometer (indoors) to watch his moving speed. However, the two methods are not sufficiently accurate. Thus, we proposed to control the walking step length for accurate localization.

4.1 Localizatoin Algorithm

In this section, we will introduce our localization algorithm. First, since we need to collect RSS from a target, we will introduce how we sample RSS. Then we present our algorithm to calculate the location of the target.

4.1.1 RSS Sampling

WiFi Signal loses power while it is in the propagation. The relationship between the distance and RSS at a receiver is presented in Formula (1).⁴⁴

$$P(d) = P(1) - 10\alpha \log(d) - W + X_\sigma, \quad (1)$$

where distance d (in meters) is the receiver-transmitter distance and power $P(d)$ is the RSS at the receiver's antenna respectively. α is the path loss exponent, W (in dB) is the wall attenuation degree, and X_σ is a normally distributed variable with mean of 0 and variance of σ^2 . X_σ is caused by phenomena including multipath. This log normal wireless propagation model is merely an approximation. In realistic settings, many factors (i.e., metal objects and multipath) can affect the propagation, making the log normal model inaccurate and hence inapplicable. The influence is especially strong in the indoor settings. Many other researchers seek to use alternative ways to model the wireless environment (i.e., recording RSS values on a set of points in the space⁴⁵).

Our RSS sampling procedure is as follows. A man with a wireless sniffer moves along a route and collects RSS samples along a route. The moving velocity is adjustable. We use the RSS samples to reconstruct the target's transmission power distribution over the route. Figure 6 shows an example of the target power distribution $S(W_d)$ over a route, and W_d is the position of the agent. Dots below the curves $S(W_d)$ represent RSS samples collected by the sniffer device. The target's orthogonal projection onto the

⁴⁴ Greg Durgin, Theodore S. Rappaport, and Hao Hu, "Radio path loss and penetration loss measurements in and around homes and trees at 5.85 GHz," *IEEE Transactions on Communications* 46, no. 11 (1998): 1484-1496; Daniel B. Faria, "Modeling Signal Attenuation in IEEE 802.11 Wireless LANs - Vol. 1," Technical Report submitted to Stanford University, Stanford, California, 2005.

⁴⁵ Andreas Haeberlen, Eliot Flannery, Andrew M. Ladd, Algis Rudys, Dan S. Wallach, and Lydia E. Kavraki, "Practical robust localization over large-scale 802.11 wireless networks," in *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking (MOBICOM)*, (2004), 70-84; Morgan Quigley, David Stavens, Adam Coates, and Sebastian Thrun, "Sub-meter indoor localization in unmodified environments with inexpensive sensors," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems 2010 (IROS10)*, Taipei, Taiwan, 2010.

route is denoted as origin O . An extreme counter example is that if the agent is running 100 meters per second and the target is transmitting 1 packet per second, we cannot reconstruct the target power space distribution $S(W_d)$ along the route because of the insufficient number of samples. In reality, it is highly possible that the packet transmission rate of a target may be quite slow. Hence, a strict control of the moving velocity is necessary.

Recall that Formula (1) gives the physical model of wireless signal attenuation. We define $S(W_d)$ as the power distribution over a route. We ignore the noise term X_σ in (1), as this does not affect the essence of our sampling theory. Furthermore, noise is of high frequency and the sampling process filters a part of the noise.

4.1.2 Localization Scheme

We use the signal sampling theory⁴⁶ to address the real problem. In reality, the built-in GPS and Accelerometer are not sufficiently accurate to indicate the moving velocity of the device. Therefore we use a human step to measure the velocity of the device.

Theorem 1: An operator holding a handheld wireless sniffer walks along a route. The RSS samples can reconstruct the target power distribution along a route in Figure 6 if and only if the space sampling interval S_I satisfies (2) and a RSS sample must be collected within each S_I ,

$$S_I < \frac{1}{2F_{\max}}, \quad (2)$$

where F_{\max} is the band of limit of $S(W_d)$, and $S(W_d)$ is the power distribution along the route in terms of distance d with respect to the original point O .

Proof: First, refer to Formula (1); we present a mathematical model of the target's power distribution $S(W_d)$ over walking distance W_d .

$$S(W_d) = P(1) - 10\alpha \log(d) - W + X_\sigma \quad (3)$$

$$d = \sqrt{(W_d - T_p)^2 + D^2} \quad (4)$$

We use Figure 7 to explain Formulas (3) and (4). Let us denote the point H in Figure 7 to be the handheld sniffer, which is also the position of the operator. W_d is the x coordinate of the operator on the route AB. T_p denotes the x coordinate of the target's projection G on the route AB. In addition, D denotes the distance between the target and its projection G. Therefore, d in the formula is the distance between the sniffer and the target. $S(W_d)$ is band limited and its cutoff frequency is denoted as F_{\max} . For example, F_{\max} can be referred to the cutoff frequency so that a large percentage (such as 95%) of the energy in the spectrum is preserved. To be able to reconstruct $S(W_d)$ from its samples, from the Nyquist sampling theorem, the sampling frequency F_s must satisfy the condition presented in Formula (5),

$$F_s > 2F_{\max}. \quad (5)$$

⁴⁶ Alan V. Oppenheim and Ronald W. Schaffer, *Discrete-Time Signal Processing*, 3rd ed., Prentice-Hall Signal Processing Series (New Jersey: Prentice Hall, 2009).

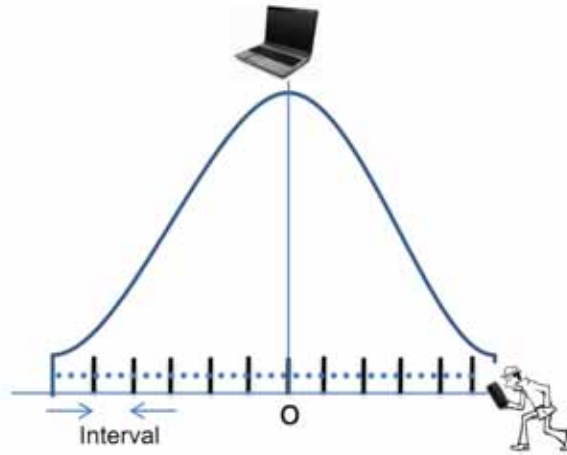
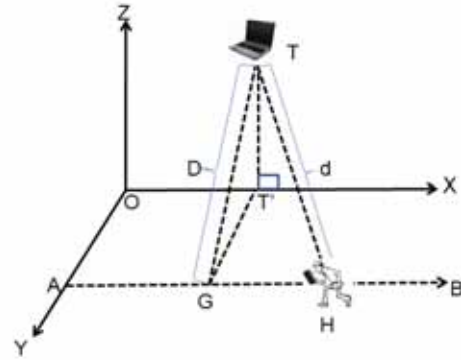
Figure 6: Power Distribution $s(t)$ over a Route.

Figure 7: Signal Strength Reading Analysis.

F_s determines how many samples we should collect in a single unit of distance (e.g., 1 meter). Accordingly, we divide a single unit of distance into F_s segments of equal distance, and we denote such distance as space sampling interval S_I . Obviously, S_I equals $\frac{1}{F_s}$. Finally, to correctly collect RSS samples, the operator should collect at least one packet within each S_I .

Theorem 1 makes localization via a hand-held walking device feasible. First, we do not need to measure walking velocity and just need to collect at least one RSS sample each S_I meters, which can be roughly measured by our step length. Second, we do not need to measure the target's packet transmission rate. We just need to wait for one RSS sample within each S_I before moving forward.

4.2 Evaluation

We have conducted real-world experiments to evaluate the performance of localization algorithm.

4.2.1 Sniffer Velocity v.s. Localization Accuracy

We placed a laptop that keeps sending out ICMP packets every two seconds in a corridor. Then, we had a robot to move along the straight route. The robot was armed with a wireless sniffer so that the robot could collect RSS samples while moving. After the robot reached the end of the route, we selected the position in the route where the robot collected the strongest signal strength as the estimated position for the laptop. In the ideal case, the x-axis of this position should equal the x-axis of the laptop's position. We set the velocity of the robot to 100mm/s, 200mm/s, 300mm/s, and 400mm/s and located the laptop. To derive the laptop's position, we used the Simultaneous localization and mapping (SLAM) function shipped with the robot to generate a map for that floor and derive the coordinates of every point in the map. We measured the difference between the laptop's x-coordinate and the x-coordinate of the estimated position. The result is shown in Figure 8. The x-axis indicates the robot's velocity and the y-axis represents the accuracy of the target laptop. This figure shows that when the velocity increases, the localization error increases.

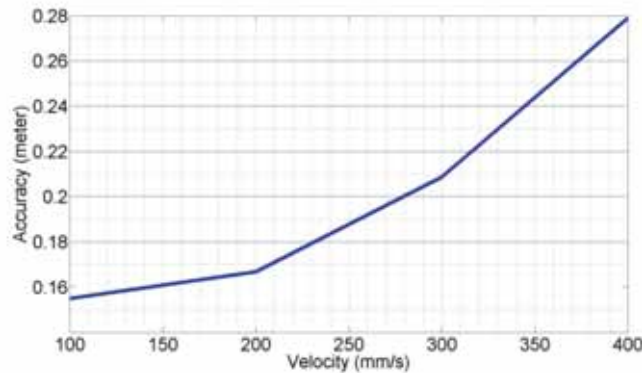


Figure 8: Sniffer Velocity vs. Localization Accuracy.

4.2.2 Failure of GPS and Accelerometer Measuring Velocity

At the very beginning, we tried to use the built-in GPS/Accelerometer to estimate the device's velocity for outdoor/indoor investigations. The GPS in N900 can obtain the velocity directly from satellites. However, the result is not accurate if the walking speed is slow.⁴⁷ We also tried to use the accelerometer to estimate the device's velocity for indoor investigations since the accelerometer reads the acceleration of the device. We simply integral the acceleration and get the velocity of the device. However, the results were again disappointing.⁴⁸ We tied N900 with a robot and controlled the robot at a stable speed. The performance of the GPS and Accelerometer is presented in Figures 9 and 10.

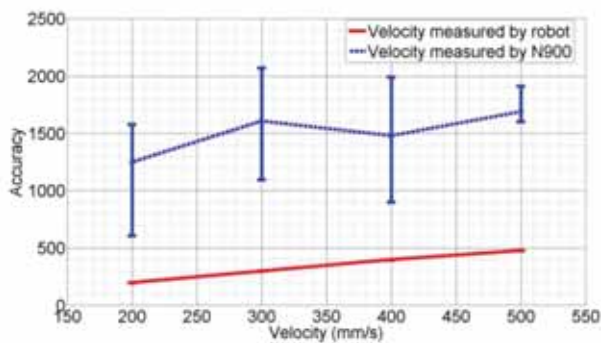


Figure 9: GPS Measured Velocity vs. Real Velocity.

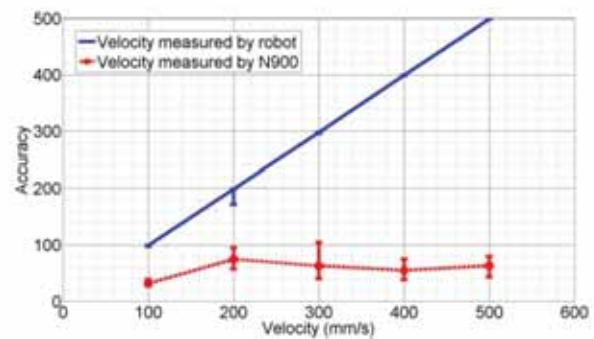


Figure 10: Accelerometer Measured Velocity vs. Real Velocity.

4.2.3 Transmission Time Interval vs. Localization Accuracy

We also conducted a set of experiments to validate the correctness of our sampling theory. We placed a laptop that keeps sending out ICMP packets as a target in a corridor and had an operator use our system to

⁴⁷ Maemo.ORG, "N900 Hardware GPS," last modified July 30, 2010, http://wiki.maemo.org/N900_Hardware_GPS.

⁴⁸ Maemo.ORG, "N900 accelerometer," last modified November 24, 2011, http://wiki.maemo.org/N900_accelerometer.

locate the laptop. The operator walked along a straight route to sniff signals transmitted from the laptop and chose the position where the strongest signal strength was sensed as the target's position.

Our evaluation contains two steps. Firstly, we analysed the relationship between the distance from the laptop to the operator's route and the length of the space sampling interval. From our analysis, we derived guidance about how long a space sampling interval should be given a specific (or estimated) distance between the operator's route and a laptop. Secondly, we utilized this result and conducted our localization evaluation using HaLo. The rest of this section will introduce the two steps in detail.

First, we focused on evaluating the length of the space sampling interval given the distance between an operator's route and a target. Recalling the experiment scenario described in Figure 7, and referring to the mathematical definition of $S(W_d)$ presented in Formula (3), we calculated the signal strength at every position along the operator's route. Then, we applied the Fourier transform to this data and identified the cutoff frequency F_{\max} . Finally, from Formula (2), we derived the value of the space sampling interval. We set the distance from the target laptop to the operator's route to 1, 2, 4, 8, 16, 32, 64 and 128 meters, and calculated the value of the space sampling interval, respectively. We presented our analysis results in Figure 11. In this figure, the x-axis represents a series of distances between the target and the operator's route, and the y-axis represents the value of the space sampling interval. As the distance increases, the length of space sampling interval increases, accordingly.

Secondly, we specifically selected the route for our operator so that the distance between the route and the target was 2 meters. According to our evaluation in the first step, the space sampling interval could be 0.1 meters, which means that in order to correctly collect the RSS samples, the operator had to collect at least one packet every 0.1 meters. The operator moved 0.1 meters at a time and sniffed the signal. To provide a clear reference for the operator in terms of how far 0.1 meters was along this route, we also had a robot beside the operator to move forward 0.1 meters at a time as a reference. In practice, the operator can be trained to know his/her step length and control his/her walking. The localization accuracy was calculated by measuring the difference along the x-axis between the estimated position and the laptop's position. We set the laptop's transmission time interval to 0.2s, 0.4s, 0.6s and 0.8s respectively, and used the N900 to locate the laptop under each setting ten times. The results are presented in Figure 12. The x-axis indicates the target's transmission time interval and the y-axis represents the accuracy of the target laptop. We can see that the mean error is close to zero, which means the algorithm is pretty good. The confidence interval is no more than 2 meters in each round test, which means the algorithm is accurate and the sampling theory is effective for real-world localization using HaLo.

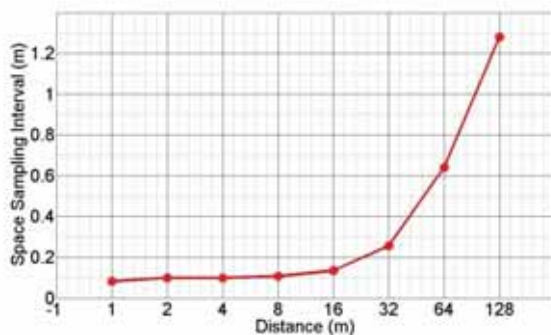


Figure 11: Space Sampling Interval vs. Estimated Target Distance.

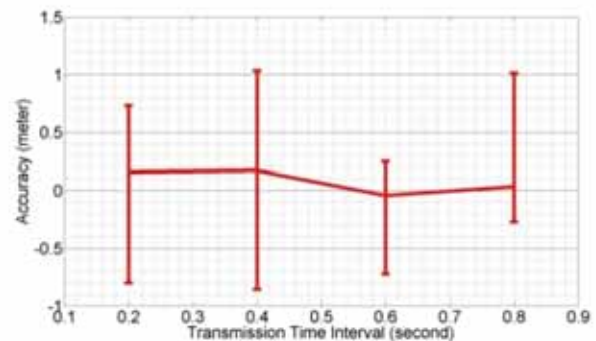


Figure 12: Transmission Time Interval vs. Localization Accuracy.

5. Conclusion

In this paper, we reviewed the current frameworks of digital forensics and found a gap between academic researchers and law enforcement in the area of network forensics. By introducing actual laws into the proposed framework, we combined academic research and actual investigation. We also developed a forensic hand-held device HaLo for law enforcement to locate suspects in real time investigation. Law enforcement can use HaLo to collect strong evidence and apply for high-level authorization such as a search warrant. We expect our refined framework can bring a fundamental guidance to network forensics research.

Implementing Digital Forensic Readiness for Cloud Computing Using Performance Monitoring Tools

F.R. Van Staden and H.S. Venter

University of Pretoria

Abstract

Cloud computing is a scalable, distributed environment that exists within the confines of the internet and is used to deliver digital services to users. The problem is that the distributed nature of cloud computing makes it difficult to collect digital forensic data after a malicious incident occurred. The article discusses using performance monitoring tools to implement digital forensic readiness for cloud computing. A learning management system is used as an example of a cloud computing implementation during the discussion.

Author

F.R. van Staden received his undergraduate degree in 2005 from the University of Pretoria and is presently studying towards his Master of Science degree at the same institution. His research interests include digital forensics, digital forensic readiness and trusted systems.

1. Introduction

Digital forensic specialists are plagued with sifting through large data sets to find incident information. As part of the collection process, system log files are collected and scanned for data about the incident. The problem is that, due to storage constraints on live systems, most log files are rotated (i.e., old data is overwritten with new data) in a bid to save space. The process of log file rotation can cause the data, pertaining to a possible incident, to be lost. Therefore, when an incident is detected, the production systems involved need to be stopped until the data collection process is complete.

Cloud computing is a scalable, distributed environment, which exists within the confines of the internet and is used to deliver digital services to users. Services might be hosted in the same physical location or the same service might be hosted in multiple physical locations. The distributed nature of cloud computing adds to the complexity of collecting digital forensic data after a malicious incident occurred. The addition of digital forensic readiness to cloud computing allows for the collection of live digital forensic data while users are accessing services.

Proper application management includes application performance monitoring. The only way to ensure that users are experiencing services as is intended by the service provider is to monitor the performance of the service. Users may assume that, since service providers guarantee the quality of service that is provided to users, performance monitoring tools are already implemented. Can a performance monitoring tool be used to implement digital forensic readiness to enhance digital forensic investigations surrounding cloud computing implementations?

Using a performance monitoring tool to collect log file data means that no time is spent, during the digital forensic investigation, to collect the log file data. Collecting live data using a performance monitoring tool means that downtime is not required for data collection purposes. Collecting live data using a performance monitoring tool implies that the collected data must be validated and stored in a read-only data environment.

This article discusses using performance monitoring tools to implement digital forensic readiness for a learning management system. Learning management systems (LMS) provide services supporting e-learning activities such as communication, document management, assessment and content management tools. Using a web browser, users can access LMS services from anywhere. The LMS could be implemented as a cloud computing service.

The rest of the article is structured as follows: Section 2 provides the background for this article; Section 3 discusses where the data is collected from and how the data is collected in a digital forensically sound manner. Section 4 discusses the basic investigations that were performed with the data. The article closes with the conclusion that it is possible to use a performance monitoring tool to collect digital forensic data from cloud computing implementations.

2. Background

In this section an overview is given of performance monitoring tools, digital forensics, digital forensic readiness, cloud computing and learning management systems. Performance monitoring tools are discussed to show the similarities that exist between the various data monitoring tools and to describe the components that are relied upon to implement the proposed solution. The digital forensics section discusses digital forensics and a digital forensic process model. Digital forensic readiness is discussed to show what is needed to make data ready to be used in a digital forensic investigation. A short overview of cloud computing and learning management systems is given to set the background for the environment that was used in the experiment.

2.1 Performance monitoring tools

Performance monitoring tools are designed to collect data about software systems to report on performance, uptime and availability. Live data about the monitored systems is used to detect system problems and pinpoint the source of the problem. Some performance monitoring tools make use of the Simple Network Monitoring Protocol (SNMP) that is specified in RFC5590.¹ SNMP specifies a method for connecting to and collecting data from servers, firewalls and network devices, mainly for the purpose of performance monitoring.

Data used by performance monitoring tools are categorised into live data, historical data and custom data. Live data is data that was collected during the latest completed collection cycle and is used to display the current system state. Live data can be kept for a period of time to show the changes in system state over time. After a set period of time the live data is reformatted and moved to the historical data set. Historical data is used to calculate the performance, uptime and availability of systems.

Custom data is data collected by the performance monitoring system but not used for displaying system state or reformatted to become historical data. Performance monitoring systems normally do not understand the meaning of custom data. Custom data is stored in read-only data tables to be used by custom reports. Descriptors can be created for custom data, to explain the meaning of the data to the performance monitoring tool, but then the data is not seen as custom anymore.

¹ D. Harrington and J. Shoenwaelder, "Transport Subsystem for the Simple Network Management Protocol (SNMP)," RFC5590. s.l. : IETF, 2009.

The performance monitoring tool mentioned in this paper uses SSH to establish communication between the probes and the performance monitoring server. According to RFC4251,² Secure Shell (SSH) is used to implement secure communication over an insecure network. The implication is that the integrity of the data sent to the performance monitor by the probes, is assured. Data integrity must be assured if the data is to be used as part of a digital forensic investigation. Digital forensics and digital forensic readiness is discussed in the following section.

2.2 Digital forensics and digital forensic readiness

Digital forensic science is a relatively new field of study that evolved from forensic science to allow crime scene investigation of digital crime scenes. According to the Oxford Dictionary,³ digital forensic science is the systematic gathering of information about electronic devices that can be used in a court of law. Digital forensic science is more popularly called digital forensics and sometimes also called computer forensics.

Palmer⁴ defines digital forensics as “the use of scientifically derived proven methods towards the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitation or furthering the reconstruction of events.” Palmer’s definition describes the digital forensic process whereas Oxford describes digital forensic science. The Digital Forensic Process Model (DFPM) by Kohn, et al.,⁵ states that; “any digital forensic process must have an outcome that is acceptable by law.”

Rowlingson⁶ defines digital forensic readiness as consisting of two objectives. The first objective is to maximise the environment’s capability of collecting digital forensic information and the second objective is to minimize the cost of a forensic investigation. Preparing any environment to be digital forensically ready, a mechanism will need to be added to preserve, collect and validate the information contained in the environment. The information gathered from the environment can then be used as part of a digital forensic investigation. Cloud computing architecture and implementation models are discussed in the following section.

2.3 Cloud computing

Cloud computing is the new buzz word in digital service delivery. According to Kaufman⁷ cloud computing is the ability to utilize scalable, distributed computing environments within the confines of the

² T. Ylonen, “The Secure Shell (SSH) Protocol Architecture,” *RFC4251*. s.l. : Network working group, 2006. RFC4251.

³ Oxford. AskOxford.com. AskOxford.com. s.l. : Oxford University Press, 2010.

⁴ G. L. Palmer, “Road Map for Digital Forensic Research. Road Map for Digital Forensic Research.” [Electronic Publication]. s.l. : Digital Forensic Research Workshop (DFRWS), 2002.

⁵ Michael Köhn, J. H. P. Eloff, and M. S. Olivier, “UML Modelling of Digital Forensic Process Models (DFPMs),” in *Proceedings of the ISSA 2008 Innovative Minds Conference, Johannesburg, South Africa, July 2008*, ed. H. S. Venter et al. (Published electronically).

⁶ R. Rowlingson, “A Ten Step Process for Forensic Readiness,” *International Journal of Digital Evidence* 2, no. 3 (2004): 1-28.

⁷ Lori M. Kaufman, “Data Security in the world of cloud computing,” *Security & Privacy, IEEE* 7, no. 4 (July-August 2009): 61-64.

internet. Spring⁸ explains that cloud computing has three distinct service models namely; Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Cloud computing can then be defined as a group name for digital services that are made available over the internet to users.

The architectural layers of cloud computing is described by Spring as being Facility, Network, Hardware, OS, Middleware, Application and User. Each cloud computing service model defines at which layer the user ownership ends and the service provider ownership begins. The service provider always controls the facility, network and hardware of the cloud computing system.

When implementing the IaaS model the user owns the Middleware and Application architectural layers and could opt to control the OS layer too. Implementing the PaaS model the user owns the Application architectural layer and could opt to control the Middleware layer. In a SaaS model implementation the service provider owns every architecture layer except for the User layer. The user is allowed to make use of a service but has no ownership or control of the service.

According to Spring,⁹ the different architectural layers should be monitored for performance and to detect security breaches. The responsibility for monitoring each architectural layer is dependent on the cloud computing model that is implemented and should be clearly defined in the SLA an organization enters into with a service provider. Part of the SLA should also stipulate what type of reports will be made available to the organisation and at what time. Organisations must be able to trust service providers with the organisational data that is stored and processed in the cloud. Data protection remains the responsibility of the organization.

Song¹⁰ suggested implementing Data Protection as a Service (DPaaS) in combination with any other cloud computing model. DPaaS makes use of different security strategies to ensure the security of data while still enabling rapid development and maintenance. Accountability for data processing is provided, with DPaaS, by implementing logging and auditing functions. Learning management systems are discussed in the following section.

2.4 Learning Management System

Learning management systems are used to manage user learning activities using a web interface. According to McGill¹¹ a Learning Management System(LMS) is used to support e-learning activities by processing, storing and disseminating educational materials and supporting administration and communication associated with teaching and learning. The LMS manages course information, online assessments, online assignments, course grades and course communications. Access to the LMS is gained by use of a web browser and an internet connection. Users of an LMS are students and lecturers that have courses hosted on the LMS.

An LMS is a collection of services that are accessed by users who take part in e-learning activities. Access to the different services offered by the LMS is controlled on course level. Lecturers can decide

⁸ Jonathan Spring, "Monitoring Cloud Computing by Layer, Part 1," *Security & Privacy, IEEE* 9, no. 2 (March-April 2011): 66-68.

⁹ Jonathan Spring, "Monitoring Cloud Computing by Layer, Part 2," *Security & Privacy, IEEE* 9, no. 3 (May-June 2011): 52-55.

¹⁰ Dawn Song, Elaine Shi, Ian Fischer, and Umesh Shankar, "Cloud data protection for the masses," *Computer* 45, no. 1 (January 2012): 39-45.

¹¹ Tabya J. McGill and Jane E. Klobas, "A task-technology fit view of learning management system impact," *Computers and Education* 52, no. 2 (2009): 496-508.

what services there students can make use of to support the students learning activities. The following section discusses the implementation of the LMS as a cloud computing service and how the LMS is monitored.

3. Implementation of the Learning Management System

LMS environments need a lot of computing resources and high system availability to provide efficient and effective services to users. Ensuring quality of service is made possible by virtualising the application server installations and combining multiple virtual servers to make up the LMS environment. The virtual server manager implements high availability by distributing virtual servers of the same type over different physical servers in different data centres. Virtualising the LMS environment in such a way is equal to implementing the LMS as a cloud computing service.

Implementing a LMS in the Software as a Service(SaaS) cloud computing model, the LMS is hosted outside the organization. The organisation owns the data that is contained inside the LMS but has no other control over the LMS. Performance monitoring should be implemented by the service provider to prove that service level agreements are met. Security monitoring and digital forensic readiness surrounding the LMS should also be included in the Service Level Agreement between the organisation and the service provider. User activity data is normally supplied by an internal LMS process. Since the organisation does not administrate the service, the user activity must be requested from the service provider.

The LMS could also be implemented using the cloud computing service model, Platform, as a Service (PaaS). The organisation cannot implement a performance monitoring tool in the PaaS model, since the service provider has ownership of the OS architectural layer. Performance and security monitoring must still be specified in the service level agreement. The organisation has ownership of the application and can access user activity data supplied by the application.

The decision was made to implement the LMS using the IaaS model and the organisation has opted to control the OS architectural layer. As part of the operational environment of the LMS, the organisation implemented a performance monitoring tool to collect data about the performance of the LMS environment. The performance data is used to ensure that users experience a good quality service. Performance data is also used to scale the LMS environment according to user utilisation figures. Figure 1 shows an example implementation of an LMS environment.

The LMS architecture is monitored by the performance monitoring system comprising of a monitoring server and a set of probes that were installed on each component in the LMS environment. Different types of monitoring probes exist, to monitor different data collection points like system performance, database tables, log files and selected system files. Probes do not normally perform any processing but connect to a data point and periodically reads the value of the data point. A data point can be a system value, also called a standard data point or it could be a software generated data point, also called an extended data point. Standard data points like CPU, memory usage and drive space usage exist as part of the operating system. Extended data points are small applications that can be created to capture and expose data from log files and databases.

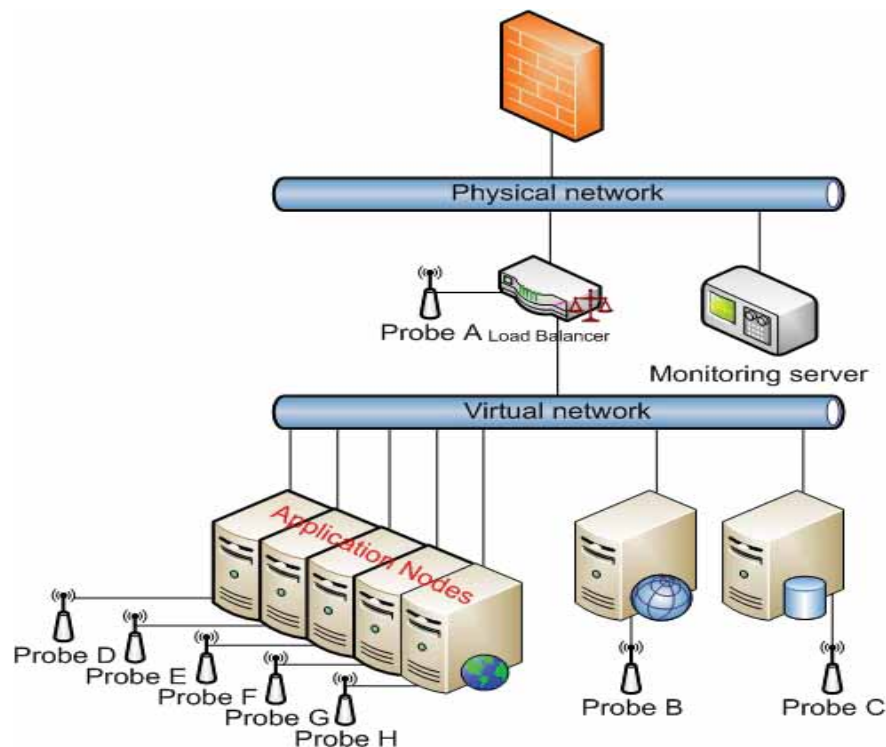


Figure 1. LMS Layout with Probes.

Communication between the monitoring probe and the performance monitoring server is established periodically, according to the timed events as set by the system administrator in the performance monitoring server. If, for example, the timed event is set to five minutes, then the performance monitoring tool will communicate with the probes every five minutes. Communication between the probe and the monitoring server is achieved using SSH or HTTPS, depending on the implementation of the monitoring tool, which means that no tampering can occur during transmission of the data and the origin of the data can be validated.

Implementing digital forensic readiness required the creation of extension points to read data from log files on all the LMS components and the central database. The extension points were defined in the performance monitoring server database using data extensions provided by the performance monitoring system. Data read from the extension points is stored in a read-only table on the performance monitoring server database. A read-only database table only allows the data to be stored using SQL INSERT statement and the data to be read, using the SQL SELECT statement. The read-only table does not allow the data to be edited using an SQL UPDATE statement or deleted using an SQL DELETE statement. The following sections discuss the function and monitoring points of the different components, as depicted in Figure 1, starting with the load balancer.

3.1 The load balancer

Users access LMS services through a load balancer device. A load balancer device does exactly what the name implies, i.e., to distribute user session load over server resources in a balanced way. Servers that are placed in a load balanced group are also placed in a virtual network. Any application server can freely

communicate with any other application server in the same virtual network but can only communicate with servers outside the virtual network by using the load balancer as a gateway device. The load balancer can also act as a firewall by blocking traffic to certain ports, and as a reverse proxy, by hiding the architecture of the load balanced servers, from users.

SSL offloading is implemented on the load balancer to enforce https between the user's browser and the LMS. Figure 1 depicts probe A being connected to the load balancer device. Probe A was extended to read data from the user session, specifically the origin address and the session_id. Since SSL offloading is implemented on the load balancer, the origin address of the user is used to verify the user's session and create a session cookie. The origin address and session_id is stored in load balancer read-only table, called usersession_from_LB, on the performance monitoring system database with a time stamp of when the session was created, as depicted in Figure 2.

user_session_from_LB	
PK	<u>Date_time</u>
PK	<u>session_id</u>
	originating_address

Figure 2. Load balancer database table.

When the initial connection is made, the originating address is used to create the session_id in the session cookie. All of the connections made to the load balancer from the same origin areas signed the same session cookie until the session times out. The session cookie contains information that allows the load balancer to route user requests to the same application node every time a user request is received during the life-time of the session. The Load balancer and the LMS share the session_id to ensure that the same user session is assigned to the application node. The following section discusses the application nodes.

3.2 The application nodes

Although Figure 1 shows all the application nodes in one location, this might not be true. As stated before, the virtual application nodes might reside on different physical servers in different data centres. The application nodes should be setup to work independently meaning that the application nodes do not know of each other but can perform the same functions. Since the LMS is web based, each virtual application node has a web server and the same collection of web applications.

Figure 1 shows probe D to probe H connected to the different virtual application nodes. Probes D to H were extended to collect data from the virtual application node's log files. Data can be collected from several log files that exist on each virtual application node. The data that was selected to be collected was data that pertains to pages that a user visits within the LMS. This data was stored on the performance monitor database in a read-only table called user_session_from_AN, with a time stamp for each record. Figure 3 depicts the user_session_from_AN table.

Each application node should still have access to the same application data, therefore a central database and a central content store is used by all the virtual application nodes. The central database is discussed in the following section.

user_session_from_AN	
PK	<u>Date time</u>
PK	<u>username</u>
	page_name browser application_node

Figure 3. Table used to store data from application nodes.

3.3 The central database

The central data store not only stores application data but also user session information. User session information is kept on the database as long as the user session is active, to ensure that a user session can be recreated in the event that one of the application nodes fail. When the user session expires the data is marked to be removed from the database by the database garbage collection process. User activity is also stored on the database in user tracer tables which is only removed when a user is deleted from the user table. User activity data is used by the LMS to profile user activity and equate the user activity with a user's academic performance. Probe C in Figure 1 was extended to collect user session data and user tracing data and save it to read-only tables, as depicted in Figure 4, on the performance monitor database. User session data is extracted daily, before the garbage collector clears the data, by querying the user session table, and stored in the user_session_from_DB table. The data is stored on the performance monitor database to ensure that the data is not lost and a link between the username and the user's originating address can be kept. Tracer data is extracted periodically by querying the user tracer tables and stored in the user_activity_from the_DB table. The central content store is discussed in the next section.

user_session_from_DB	
PK	<u>Date time</u>
PK	<u>username</u>
	session_id

user_activity_from_DB	
PK	<u>Date time</u>
PK	<u>username</u>
	action_performed

Figure 4. Tables used to store the data obtained from the central database.

3.4 The content store

LMS content, like document files or movie files are stored in the central content store. Access to LMS content is controlled by the roles that are defined in the LMS. An example would be that, if content is allocated to a specific course in the LMS, only users enrolled in that course will be able to access the content.

Actions initiated for content items are create, allocate, access, de-allocate and delete, are all stored in the content log file. Probe B was extended to collect the action data from the content log file. To safeguard against the accidental deletion of content items, a process was implemented on the content store

that archives deleted content. The archive process creates a log file entry in the archive log. Probe B was extended to store the archive log entry. Figure 5 depicts the `user_activity_from_CS` table, that is used to store data from the content log file, and the `archive_action_status_CS` table, that is used to store the data from the archive log file.

user_activity_from_CS		archive_action_status_CS	
PK	<u>Date time</u>	PK	<u>Date time</u>
PK	<u>content item name</u>	PK	<u>content item name</u>
	session_identifyer		archive_status
	action_performed		

Figure 5. Tables used to store data from central content store log files.

Section 4 gives an overview of the data collected and the way the data could be used.

4. Collecting digital forensic data from the LMS

This section discusses the data that was collected from the different probes and how the data was used to support different investigations. Firstly, we should prove that the data could be used for a digital forensic investigation. To support a digital forensic investigation the data must be collected, preserved and validated, using scientifically proven methods.

The data was collected using performance monitoring probes that were extended to collect the data from system sources that would normally be used in an investigation. Communication between the probes and the monitoring server was achieved using SSH, which means that no tampering could occur during transmission of the data and the origin of the data can be validated.

Preservation of the collected data was implemented by extending the performance monitoring system's database with extra read-only database tables. As stated before, the tables allow for the addition of more data and the reading of the stored data. Read-only database tables do not allow stored data to be changed or deleted. The process of storing the data makes use of xml files to specify the data mapping between the data the probes sent and to specify where in the database the data is stored. Three experimental investigations were constructed to show that the data could be used to support an investigation. The following sections discuss the experimental investigations that were attempted and the outcome of the investigations.

4.1 Tracing the origin of a user

The first question posed was, using the data we collected, could we determine the origin of a specific user? A database view called `user_origin` was created using the tables `user_session_from_LB` and `user_session_from_DB`. Querying the view with a username produced a list with all the sessions and the origins that were logged for the user. Since the user could have multiple sessions that showed different origins over time, the list was sorted according to the stored time stamps.

To prove that the data was correct, a test was done using ten different users over a period of one month. The test users recorded the times that they accessed the system and the address of the connecting

machines. From the comparison of the data logged by the users and the data stored on the database it was proven that the data in the list was accurate. This proved that we could, using the data collected, determine the origin of a specific user as long as we also knew the time that the user accessed the system.

An added result of analysing the list was that we detected times where the same user accessed the LMS at the same time but from different origins. Users that were identified as accessing the LMS from different origins at the same time were forced to change their system passwords. The following section discusses using the data collected to reconstruct user activity.

4.2 Reconstructing user activity

Another question was; using the data we collected, could we reconstruct a user's activity in the system over a period of time? A database view called `user_activity` was created using the tables `user_session_from_DB`, `user_activity_from_DB` and `user_session_from_AN`. Querying the view with a specific username a list was generated that showed all the actions that a user performed during the sessions a user was logged in. Since users could performed different actions in different sessions over time, the time stamp and `session_id` was user to sort the list.

The same ten users from the previous section were used to prove that the data in the list was correct. The users were asked to record their activity on the LMS over a month. Correlating the activity recorded by the users and the data in the list it was proven that the data in the list was correct. Using the data in the list it was possible to reconstruct a user's activity on the LMS for a specific time period. The following section discusses using the data collected to determine when content was deleted and by whom it was deleted.

4.3 Determining when and by whom a content item was changed

And finally, the question was posed that, using the data we collected, could we determine when and by whom a content item was changed. A view called `user_content_activity` was created using the tables `user_session_from_DB`, `user_activity_from_DB`, `user_session_from_AN` and `user_activity_from_CS`. Data that did not pertain to the content store activity was excluded from the view. Querying the view, using the name of a specific content item, and sorting the result using the time-stamp, produced a list of user actions performed on the content item through time. The list was further reduced by excluding access events.

To prove that the data collected was correct a set of content items were created, randomly changed and then all the content items were deleted by different users. The users that took part in the testing were asked to record the actions that they performed in the content items and the time that they performed them. Correlating the data the users collected and the data in the list proved that the data in the list was correct. Using the final list, it was possible to determine when and by whom a content item was created, changed or deleted.

5. Conclusion

The paper posed the question; can a performance monitoring tool be used to implement digital forensic readiness in cloud computing in order to enhance digital forensic investigations? A performance monitoring tool extended to collect data from a Learning Management System. The Learning

Management System was used as an example of a cloud computing implementation. Experimental digital forensic investigations were performed with the data that was collected by the performance monitoring tool.

It was proven that data can be collected, preserved and validated using a performance monitoring tool. The data was collected from the live system in real-time which meant that no data was lost due to log file rotations or database garbage collation processes. Investigations could be performed without any costly downtime on the LMS since the data was preserved on a different system to the LMS. Investigators need not collect data by sifting through log files on the LMS which also speeds up the investigation process.

Experimental investigations were successfully performed using the data collected which proved that the data could be used for investigations. Investigations were completed by running simple queries on the performance monitoring database tables. Therefore it was proven that a performance monitoring tool could be used to implement digital forensic readiness in cloud computing in order to enhance digital forensic investigations.

6. Future work

After collecting data from the probes for a time it was found that the data storage requirement for implementing digital forensic readiness might at some point in the future become too expensive to be sustainable. Future work would include implementing automated processes to detect possible incidents and only collecting data for those possible incidents to reduce the amount of data collected.

Acknowledgment

This work is based on research supported by the National Research Foundation of South Africa (NRF) as part of a SA/Germany Research cooperation programme. Any opinion, findings and conclusions or recommendations expressed in this material are those of the author(s) and therefore the NRF does not accept any liability in regard thereto.

Security Monitoring for Wireless Network Forensics (SMoWF)

Yongjie Cai and Ping Ji

Abstract

With the broad deployment of WiFi networks nowadays, it is easy for malicious network users to camouflage their true identities through randomly hopping onto open wireless networks, conduct an attack and leave without being caught. Most of the current infrastructures of wireless networks do not keep logs of network activities by default, which makes it hard to obtain important network traces that may facilitate future forensics investigations for a suspicious network event. In this paper, we outline a Security Monitoring System for Wireless Network Forensics (SMoWF), which aims to establish a forensic database based on encrypted (or hashed) wireless trace digests, and to answer the critical investigation question: which wireless device appeared at where during what time? We propose to accomplish our goal through three steps: 1. Design a network trace logging method that records the abstract of useful fields of network packets. Here only abstracts of packets are kept due to privacy protection concerns. 2. Design a query/search system that allows users to conduct forensic analysis based on gathered traces; 3. Study and integrate localization algorithms into SMoWF, which can provide the location estimation of a given device when such information is needed.

Authors

Yongjie Cai is a second-year student of the Computer Science Ph.D. Program of the Graduate Center of City University of New York (CUNY). Prior to joining the graduate program of CUNY, Yongjie obtained her B.E degree in Computer Science from NanKai University, China, in year 2010. Working in the Computer Network and Mobile System Security (NeMo) Lab led by Prof. Ping Ji, Yongjie's current research interest includes Wireless Network Performance and Security Measurement, Network Traffic Analysis and Wireless Network Applications.

Ping Ji is a Professor of the Mathematics and Computer Science Department of John Jay College of Criminal Justice of the City University of New York, and a faculty member of the Computer Science Ph.D. Program of CUNY – Graduate Center. Prof. Ji holds a Ph.D. degree in Computer Science from the University of Massachusetts at Amherst and a B.S. degree in Computer Science and Technology from Tsinghua University, China. Prof. Ji's research interests cover the broad area of Computer Networks and have recently focused on Wireless and Mobile Network Security and Forensics. Since joining the City University of New York from year 2003, Prof. Ji has been active in both research and teaching, and awarded a number of CUNY internal and external government grants. These awards and grants include the RF-CUNY research award for consecutive five years, US/UK Army research grants on sensor network information quality study, and NSF grants in the field of Network Forensics.

1. Introduction

Cybercrime is an exploding security challenge in the current digital age, and has been largely concerned by the public over the past several decades. With the escalating deployment of WiFi networks, the accelerated usage of mobile devices, and the dynamic physical and protocol characteristics of wireless communication, wireless links have become an increasingly popular channel for cyber criminals to camouflage their true identities. For example, a hacker may drive on the street, randomly pick an open WiFi network, conveniently connect to the Access Point, upload or download malicious files through the Access Point, then close the session and drive away. The whole process may only take minutes to

accomplish, and when the victim machine notices the attack, the best *point of interest* that it can trace back is very likely only the benign Access Point, through which the true attacker conducted the malicious activity. It is almost always certain that the hacker will be cut loose.

In this research, we propose to design a distributed Security Monitoring system for Wireless network Forensics (SMoWF), which monitors Wireless LAN activities. *Abstracts* of network traces are captured and selectively recorded at each monitoring point. Distributed monitoring points collaborate to reconstruct the *crime scene* based on monitored logs, and the SMoWF system should be able to answer the following questions: 1. Was a particular wireless device involved in a given malicious network activity? 2. Can this device be uniquely identified by the logs? 3. Where a particular device was physically located during a given period of time.

We propose to accomplish our goal through three steps: 1. Design a network trace logging method that records the *abstracts* of useful fields of network packets. Here only *abstracts* of packets are kept for privacy protection purpose. 2. Design a query/search system that allows users to conduct forensic analysis based on monitored traces. 3. Study and implement *localization* algorithms that can provide the location information of a given device when necessary.

The rest of this paper is organized as follows: Section 2 explores the related work. Section 3 outlines the architecture of SMoWF. Section 4 illustrates wireless network trace capturing and preprocessing methods. Section 5 discusses the approaches to store critical logs and conducts post analysis and investigation. Section 6 shows the prototype of SMoWF. Section 7 concludes the paper.

2. Related Work

There are a number of wireless traffic capturing tools¹ including Wireshark, Tcpdump and Kismet/KisMac, with which we can gather wireless network traces through “off-the-shelf” 802.11 network cards. All traffic in the same network can be passively captured when a network card is set in promiscuous (i.e., monitor) mode. When the card is in monitor mode, no packets are transmitted through it and all the traffic in a specific channel can be preserved into a backend server. More interestingly, Kismet² is able to hop channels to cover the entire spectrum, and record the physical location of a monitoring point when the tool is used with a GPS receiver. Important trace information, such as the SSID, channel number, MAC address and associated clients of wireless networks in range, can be gathered by these traffic capturing tools, which may contain vital clues for future forensic investigations.

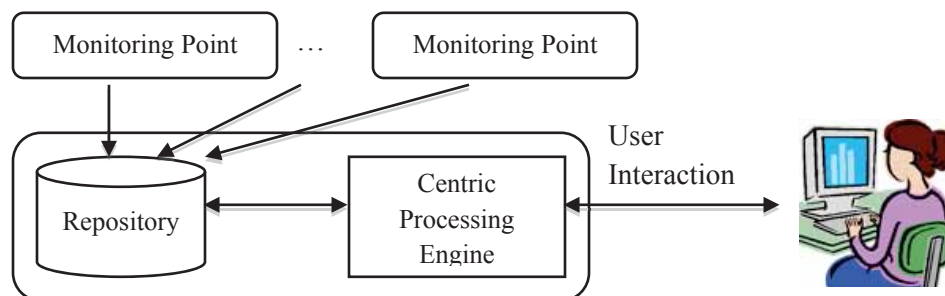


Figure 1. A typical wireless monitoring system

¹ Wireless sniffer: <https://personaltelco.net/wiki/wirelessniffer>.

² Kismet: <http://www.kismetwireless.net/>.

Researchers have proposed several wireless monitoring infrastructure systems, primarily for improving wireless channel and protocol performance. The framework of a typical wireless monitoring system is showed in Figure 1, which consists of three parts: the monitoring point, the data repository and the centric processing engine. Each monitoring point gathers network information from access points or by capturing traffic in the air, and transmits the gathered raw data to the repository. The centric processing engine conducts network analysis and reports abnormal events to network operators. VISUM³ delegates the monitoring task to a set of distributed agents using SNMP. It uses device-specific XML profiles to map retrieved high-level monitoring information to device-specific SNMP Object Identifiers. The centric processing engine is responsible to assign the subset of network devices that needs to be monitored to individual agents. This is a complicated task when the number of devices gets very large, and things can get worse when we don't know the locations of these devices.

Also along this line of research, DAIR⁴ is a framework that manages and troubleshoots enterprise wireless networks using desktop infrastructure. It proposes to attach USB-based wireless adapters to desktop machines that usually have spare CPU, disk resources and the more reliable wired-line Internet connectivity. These inexpensive adapters then work as monitoring points and can be densely deployed to cover an entire local area. In addition, Jigsaw⁵ deploys 192 stand-alone radio sniffers to monitor a wireless network that consists of 40 open APs, which cover four floors and the basement in a building. The three aforementioned systems are designed for network administrators to better monitor and diagnose the network performance of 802.11 networks. They mainly focus on maintaining the stability of clients' connectivity, reducing the interference and packet delay. The infrastructures of DAIR and Jigsaw can be adopted in our wireless network security monitoring project for raw data collection in indoor environment.

Similar to what we hope to propose, FLUX⁶ is a prototype of forensic monitoring system based on CoMo platform.⁷ It aims to identify suspicious activities, network anomalies and provide incident playback. This work proposes a similar goal with ours, however FLUX was in its preliminary stage and seemed discontinued.

Another aspect related to our work is device identification. Malicious attackers can easily camouflage their device IDs. MAC spoofing is a perfect example here for simple and effective anonymity tactics. Attackers can change the MAC address of their devices easily. However, in recent years, researchers have proposed quite a few ways to fight against MAC spoofing. First, Jeffery et al.⁸ demonstrate that with 90% accuracy 64% of users can be identified *without* using MAC address. The

³ Camden C. Ho, Krishna N. Ramachandran, Kevin C. Almeroth, and Elizabeth M. Belding-Royer, "A scalable framework for wireless network monitoring," in *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, WMASH '04 (New York, NY: ACM, 2004), 93–101.

⁴ Paramvir Bahl, Jitendra Padhye, Lenin Ravindranath, Manpreet Singh, Alec Wolman, and Brian Zill. Dair, "A framework for managing enterprise wireless networks using desktop infrastructure," in *HOTNETS'05*, 2005.

⁵ Yu-Chung Cheng, John Bellardo, Péter Benkő, Alex C. Snoeren, Geoffrey M. Voelker, and Stefan Savage. Jigsaw: solving the puzzle of enterprise 802.11 analysis. *SIGCOMM Computer Communication Review* 36 (August 2006): 39–50.

⁶ Kevin P. McGrath and John Nelson, "FLUX: A Forensic Time Machine for Wireless Networks," in *INFOCOM Poster and Demo Session*. IEEE, April 2006

⁷ Gianluca Iannaccone, "Como: An open infrastructure for network monitoring – research agenda," *Intel Research Technical Report* (February 2005).

⁸ Jeffrey Pang, Ben Greenstein, Ramakrishna Gummadi, Srinivasan Seshan, and David Wetherall, "802.11 user fingerprinting," in *Proceedings of the 13th annual ACM international conference on Mobile computing and networking (MobiCom '07)* (New York, NY: ACM, 2007), 99–110.

implicit identifiers from users' network activities, such as pairs of IP Address and port, SSID probes, broadcast packets sizes and MAC Protocol Fields, can help identify a unique user (device) quite accurately. S. Dolatshahi et al.^{9, 10} show the effectiveness of using RF signature as a wireless device identity. They exploited the imperfection of commercially used RF transmitter and amplifiers, which is difficult to for attackers to modify. Moreover, Polak et al.¹¹ propose a method by the analysis of the in-band distortion and the spectral growth to uncover the more sophisticated attackers who distorted their signatures. In the current stage of our work, we use MAC address as the device identifier without worrying too much about MAC spoofing problem. However, in the future deployment of SMoWF system, we will consider the above mentioned device identification methods and implement appropriate ones to fight against MAC spoofing.

3. Overview of SMoWF

The emerging and increasing growth of WiFi wireless networking technology makes it possible to connect to the Internet from anywhere at anytime. For example, in a wireless network measurement study,¹² we conducted experiments around a three-block metropolitan neighbourhood of the mid-west side of Manhattan for 12 runs, and detected 8000+ access points deployed in the neighborhood. The densely deployed WiFi networks are undoubtedly making our life much easier and enjoyable, but they also provide more opportunities for malicious users to conduct criminal activities through mobile devices. We notice that among our detected access points, about 30 percent provide unencrypted WiFi services. In other words, these open networks can be easily compromised.

In this paper, we propose a security monitoring infrastructure for wireless network forensics (SMoWF), which is to build an intelligent monitoring system that can uncover malicious devices, track their activities in Wireless LAN of metropolitan area, and preserve digital evidence to facility future cyber crime investigation. SMoWF system should be able to answer the following questions:

- Whether or not a particular device was involved in a given malicious network activity?
- Can this device be uniquely identified by the logs?
- Where was a particular device physically located during a given event?

Similar to Figure 1, the SMoWF system consists of a set of monitors that are responsible to capture Wireless network traffic. These monitoring points are distributed through a Wireless network and may be moved around to cover Wireless LANs as much as possible. After the collection of raw traffic data, SMoWF parses raw data into human-readable texts, eliminates irrelevant traffic types and extracts useful information for device identification and localization. It also removes the data part of traffic packets to

⁹ S. Dolatshahi, A. Polak, and D. L. Goeckel, "Identification of wireless users via power amplifier imperfections," *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*.

¹⁰ A. Polak, S. Dolatshahi and D. Goeckel, "Identifying Wireless Users via Transmitter Imperfections," *IEEE Journal on Selected Areas in Communications - Special Issue on Advances in Digital Forensics for Communications and Networking* 29, no. 7 (August 2011): 1469 - 1479.

¹¹ A. C. Polak and D. L. Goeckel, "RF Fingerprinting of Users Who Actively Mask Their Identities with Artificial Distortion," in *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference, 6-9 November 2011*, 270-274.

¹² Yongjie Cai and Ping Ji, "A measurement study for understanding wireless forensic monitoring," (to appear in *ICDFI*, Sept. 2012).

protect users' privacy. SMOWF uses a central repository to store processed data as digital evidence. Finally, it includes a post-investigation engine that helps investigators to figure out what was going on when a criminal activity occurred. The post-investigation engine retrieves relevant data from the evidence repository and is able to answer the aforementioned questions.

4. Traffic Capture and Preprocess

Comparing to those monitoring systems deployed in buildings/universities,^{13, 14} there are several challenges of Wireless traffic monitoring in a metropolitan area: 1) the number of access points that are observable is large; 2) the AP locations and distributions are unknown; 3) we are out of control of these access points. Therefore, the traditional ways of obtaining traffic via access points are not practical. We cannot configure all these access points to log their real-time traffic, nor can we deploy thousands of static stand-alone monitor nodes to cover the whole area.

For SMOWF, we propose to delegate the traffic capturing tasks to wireless monitoring points, such as laptops being either stationary or mobile. These monitoring points passively capture nearby Wireless network traffic, and periodically upload the encrypted or hashed traffic logs to a central repository. Particularly, in our experiments, we use Kismet installed on a MacBook Pro to gather raw Wireless network traffic. Kismet is an 802.11 wireless network sniffer working with any wireless card, which supports monitoring mode, and detects networks by passively collecting packets. It can provide GPS coordinates where packets are detected when integrated with a GPS device. Kismet will generate several log files including *.pcapdump*, *.gpsxml*, *.netxml*, *.nettxt*, *.alert*. All above MAC layer packets information, together with Per-Packet Information (PPI) header that includes channel, signal and noise strength, are logged to *.pcapdump* files. GPS information such as coordinates and speed are recorded into *.gpsxml* files. Our SMOWF system mainly uses these two types of logs.

While Kismet logs collect raw packets in libpcap format into *.pcapdump* files, we use Tshark,¹⁵ which is the command line version of Wireshark to parse them into human-readable text files. Also we filtered out the data part of packets and only preserve packet headers.

5. Evidence Preservation and Post-investigation

For evidence preservation, only extracting packet headers to reduce logged data size is not efficient enough. The network traffic size can be huge compared to limited storage. For instance, in Section 6, we collected 362,305 packets using Kismet, about 311MB trace data by randomly walking around a three-block neighborhood for 12 trips around four hours. These data only came from one single monitor. If tens or hundreds of monitors participate, it can easily get 10-100 GB traces in one day. For instance, Jigsaw¹⁶

¹³ Camden C. Ho, Krishna N. Ramachandran, Kevin C. Almeroth, and Elizabeth M. Belding-Royer, "A scalable framework for wireless network monitoring," in *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, WMASH '04 (New York, NY: ACM, 2004), 93–101.

¹⁴ Cheng et al., "Jigsaw: solving the puzzle of enterprise 802.11 analysis."

¹⁵ Tshark: <http://www.wireshark.org/docs/man-pages/tshark.html>

¹⁶ Yu-Chung Cheng, Mikhail Afanasyev, Patrick Verkaik, Péter Benkő, Jennifer Chiang, Alex C. Snoeren, Stefan Savage, and Geoffrey M. Voelker, "Automating cross-layer diagnosis of enterprise wireless networks," in *SIGCOMM '07* (New York, NY: ACM, 2007), 25–36.

collected 96GB 802.11 raw traces in one day using 192 radio monitors. We made statistic analysis on the packet types in 802.11 on our data.¹⁷ We observed that half of the packets were beacons sent from access points, which simply claimed their existence and were not related to their associated clients. Therefore, we filtered out this kind of packets. Secondly, to support efficient queries and conducting post forensics investigation, we store our network traces into a database.

The critical part of our system is post-investigation, which aims to answer the questions described in Section 3. As a preliminary work for our system, we use MAC addresses as the unique identifiers of mobile devices and explore the device localization problem accordingly. We study and evaluate two localization algorithms, one is *weighted centroid algorithm*¹⁸ and the other is *log-distance path loss modeling method*.¹⁹ Weighted centroid algorithm, as Equation 1 shows, estimates the location of the target device as the weighted sum of all locations where it was observed. Shown in Formula 1, \hat{p} is estimated location of target device, p is the i th location coordinate where the device is detected, and the weight w_i is proportional to signal strength received from the target device at i th location.

$$\hat{p} = \sum_i w_i p_i, (w_i \propto s_i, \sum_i w_i = 1) \quad (1)$$

The log-distance path loss modeling method describes that the average received signal strength decreases logarithmically with distance whether in outdoor or indoor radio channels, shown in Equation 2.

$$s_i = S - 10\gamma \log d_i + X_\sigma \quad (2)$$

s_i is the received signal strength from target device at position i . d_i is the physical distance from target device with coordinates $\langle x, y \rangle$ to the monitor point with coordinates $\langle x_i, y_i \rangle$. The path loss exponent γ indicates the loss rate of the received signal strength. S is the signal strength from the device at a distance of one meter. To compensate for the random shadowing effects in radio propagation, X_σ is added as a zero-mean Gaussian distributed random variable with standard deviation σ . Theoretically, we need four monitor points or traces to determine four parameters $\langle S, \gamma, x, y \rangle$ of the target device in order to know the location coordinate $\langle x, y \rangle$. However, in practice, more than four sets of traces from the target devices are collected. We have to solve a set of over-determined equations. There are several solutions to this problem. For example, Krishna²⁰ proposed to find solutions to minimize the least mean absolute error of equations. In our system, to simplify the implementation, we used trust-region-reflective optimization approach²¹ implemented in Matlab to minimize the least square error which is defined in Equation 3.

$$J = \sum_i (s_i - S + 10 \gamma \log d_i)^2 \quad (3)$$

¹⁷ Cai and Ji, "A measurement study for understanding wireless forensic monitoring."

¹⁸ Yu-Chung Cheng, Yatin Chawathe, Anthony LaMarca, and John Krumm, "Accuracy characterization for metropolitan-scale Wi-Fi localization," in *Proceedings of the 3rd international conference on Mobile systems, applications, and services*, MobiCom '05 (New York, NY: ACM, 2005), 233–245, <http://doi.acm.org/10.1145/1067170.1067195>

¹⁹ Theodore Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. (Upper Saddle River, NJ: Prentice Hall PTR, 2001).

²⁰ Krishna Chintalapudi, Anand Padmanabha Iyer, and Venkata N. Padmanabhan, "Indoor localization without the pain," in *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, MobiCom '10 (New York, NY: ACM, 2010), 173–184

²¹ lsqnonlin: <http://www.mathworks.com/help/toolbox/optim/ug/lsgnonlin.html>.

6. Experiments and System Prototype

We conduct experiments to explore the feasibility and evaluate the performance of our system in the testbed. Our testbed, shown in Figure 2, is a three-block metropolitan area of the upper-west side in NYC, which is around 260m*260m. We use a MacBook Pro laptop with internal airport wireless card and a BU353 GPS receiver as a moving monitor point. Kismet, installed on the MacBook Pro, is configured to hop on channels to cover the entire spectrum and log all received wireless packets. We walked around the testbed for 12 runs along the path of A-H or H-A in a week of April of 2011. We collected 362,305 packets around 311MB traces.



Figure 2. Testbed and Testing Path

After parsing *.pcapdump* files into readable texts, we filter out the data payload of packets, extract packet header fields and dump them into PACKET table. The fields include frame date and time, source address, destination address, BSSID, transmitter address, and receiver address of MAC, data length, channel frequency, received signal strength, noise strength, type and subtype of 802.11, source and destination address of IP, source and destination port of TCP and UDP. Notice that a packet doesn't include all the fields. For example, 802.11 Acknowledgement and Clear-To-Send packets only contain receiver MAC address and no other MAC address. One packet only has source/destination port either from TCP or UDP. We can obtain neither TCP nor UDP information from encryption packets. Furthermore, we extract *.gpsxml* files and dump them into MAC_GPS table in our database. MAC_GPS table contains date, time, source MAC address, signal, noise, latitude, longitude, altitude, fix, speed, heading. To speed queries, we create indexes on date fields in both tables. For device localization, we chose to apply the simple but effective *weighted centroid algorithm* in our system.

We further developed a simple web user interface to help investigators to trace their interested devices. As shown in Figure 5, an investigator can enter the date and time period of an interesting event, as well as the MAC address, IP, or BSSID of a device. SMOwF then pulls out the records/packets that are related to their interested device from the database. It will estimate the locations of the device every five minutes during that time window shown as the second picture and generate a KML file that tags the geo locations of the device in Google Earth, shown in the third picture. In this way, the investigators can easily locate their interested device.

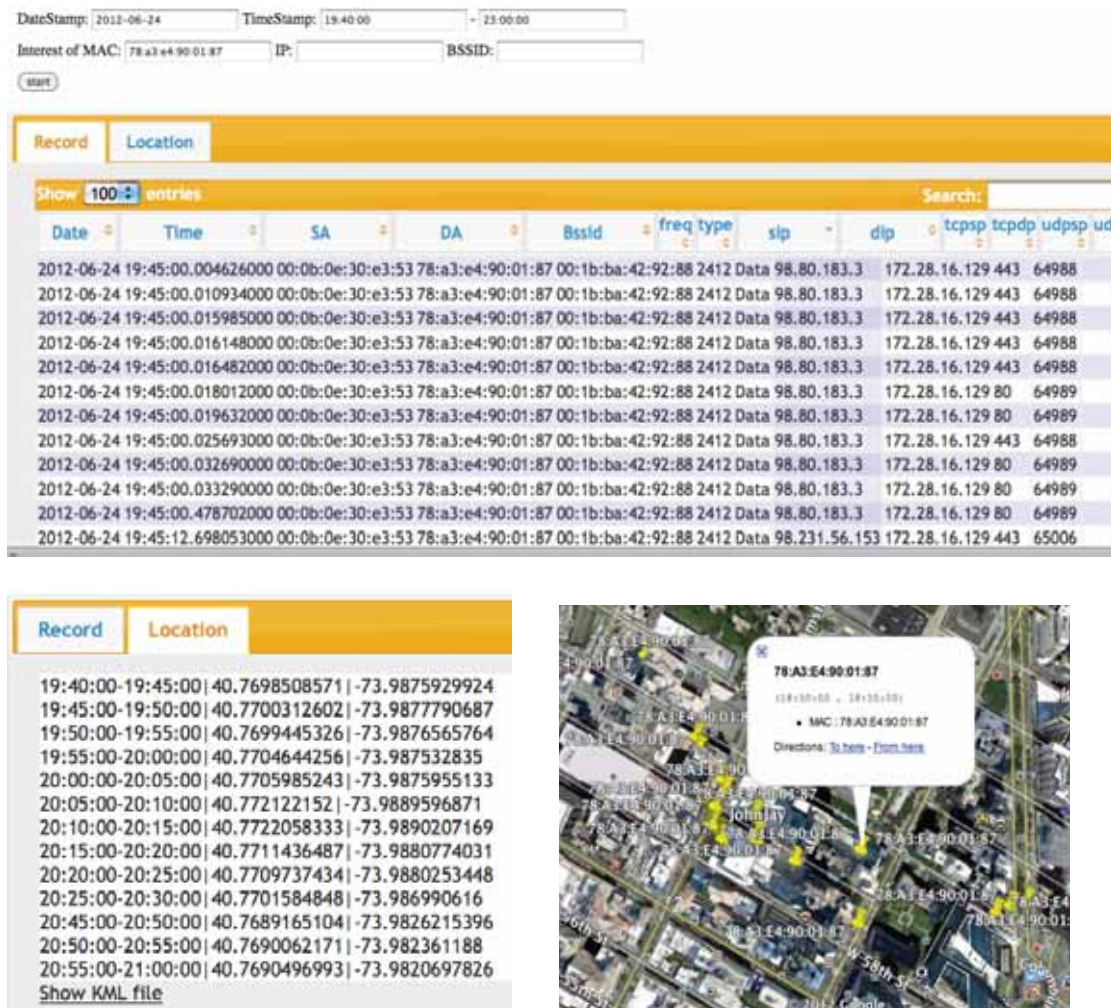


Figure 3. System Prototype

7. Conclusion

In this work, we outlined a wireless forensic monitoring system (SMoWF), which aims to establish a forensic database based on encrypted (or hashed) wireless trace digests, and to answer the following investigation questions: 1. Was a particular device involved in a given malicious network activity? 2. Can this device be uniquely identified by the logs? 3. Where a particular device was physically located during a given event. We conducted research and experiments for the following tasks: 1. Design network trace logging method that records the *abstract* of useful fields of network packets. Here only *abstracts* of packets are kept for privacy protection purpose. 2. Design a query/search system that allows users to conduct forensic analysis activities based on monitored traces. 3. Study and propose *localization* algorithms that can provide the location information of a given device.

Acknowledgments

This research is supported by the National Science Foundation under NSF grant CNS-090490

Workshops

The Archivemata Project

Meeting Digital Continuity's Technical Challenges

Peter Van Garderen,¹ P. Jordan,² T. Hooten,² C. Mumma¹ and E. McLellan¹

¹Artefactual Systems; ²International Monetary Fund

Abstract

Archivemata is a free and open-source digital preservation system developed by Artefactual Systems in part through funding from UNESCO. This interactive session will present the history of Archivemata and an overview of the system's design, features and technical architecture. It will also provide examples of how the software is being used in projects worldwide, such as the International Monetary Fund, the University of British Columbia Library and the City of Vancouver Archives. The presentation will conclude with a discussion of future directions for the system, including the development of new features; the open-source business model; training and support services; and ongoing challenges to implementing Archivemata in developing countries.

Authors

Peter Van Garderen is president of Artefactual Systems Inc., which he launched in 2001 to provide technology analysis and implementation services to archival organizations, in particular those that want to develop and implement open-source solutions. Since that time Peter has worked with a wide range of clients, providing services that range from writing strategy reports, analysing system requirements, designing technology architectures, developing software, and managing open-source projects. He is a Doctoral Candidate in Archival Science at the University of Amsterdam and a Distinguished Alumnus of the University of British Columbia's Master of Archival Studies (1997). Peter also holds a Certificate in Software Engineering (2001) from UBC. He is a regular conference speaker on the topics of digital preservation and archives technology.

Paul Jordan worked for four years as the Digital Archivist for the International Monetary Fund. Prior to that he worked as an archivist for both Occidental College and the Mary Pickford Foundation. He received his Master of Science in Information from the University of Michigan, and is currently working as the Digital Asset Administrator for Kaplan University.

Evelyn McLellan is systems analyst and consultant on Artefactual's digital preservation projects. She provides end-user support, user documentation, quality assurance testing and system requirements management for Artefactual's Archivemata project. Evelyn is a graduate of the University of British Columbia's Master of Archival Studies program (1997) and has over 10 years experience as an archivist and records information analyst. She was a co-investigator on the International Research on Permanent Authentic Records in Electronic Systems (InterPARES 3 Project) and served as an Adjunct Professor at the University of British Columbia's School of Library, Archival and Information Studies. Evelyn has written several articles and delivered a number of presentations and workshops on the topics of digital preservation and open technologies for archives.

Courtney C. Mumma is a systems analyst who provides end-user support, user documentation, quality assurance testing and system requirements management for Artefactual's Archivemata project. As a digital archivist at the City of Vancouver Archives, she helped to develop and implement their digital archives system while managing the acquisition of the hybrid digital-analogue 2010 Winter Games archives. She has been a researcher and co-investigator on the International Research on Permanent Authentic Records in Electronic Systems (InterPARES 3 Project), researcher on the UBC-SLAIS Digital

Records Forensics Project, and a member of the Professional Experts Panel on the BitCurator Project. Courtney has been published in *Archivaria* and has delivered a number of presentations and workshops nationally and internationally on the practical application of digital preservation strategies.

1. History of the Archivemata Project

In June of 2007, Kevin Bradley of the National Library of Australia with Junran Lei and Chris Blackall of the Australian Partnership for Sustainable Repositories, published “Towards an Open Source Repository and Preservation System: Recommendations on the Implementation of an Open Source Digital Archival and Preservation System and on Related Software Development” for the UNESCO Memory of the World Programme Sub-Committee on Technology. Bradley et al. advocated building sustainable systems instead of expecting some permanent storage media to solve digital preservation challenges. Their report defined open source software requirements for the implementation of a digital archival and preservation system that would consider all aspects of a digital repositories as defined by the ISO 14721 Open Archival Information System (OAIS) functional model;¹ Ingest, Access, Administration, Data Management, Preservation Planning and Archival Storage, including storage media and management software. Further, the report claimed that digital preservation solutions for simple digital objects were well understood, and that “what is needed are affordable tools, technology and training in using those systems.”²

The Sub-Committee identified existing gaps and made recommendations for the development and packaging of an Open Source Digital Preservation System. Ultimately, they concluded that what was needed was an affordable, sustainable approach that could leverage the expertise of larger institutions with more resources to innovate and share solutions with the digital preservation community at-large. Such collaborative innovation could look to the open source software development community for a model of “how a sustainable archival system might work, be sustained, be upgraded and be developed as required.”³ Most significantly, the report recommended that UNESCO support “the aggregation and development of an open source archival system, building on, and drawing together existing open source programs.”⁴

Concurrently, Artefactual Systems, Inc., was busy developing their generic Qubit⁵ information toolkit software as AtoM (Access to Memory), an open source, web-based archival description software based on International Council on Archives (ICA) standards. The UNESCO report coincided with the Artefactual team, some of their clients and the digital preservation community at large realizing there was a need for an open-source, sustainable, OAIS-based digital preservation system.

Therefore, the Archivemata project had its beginnings as the back-end digital preservation system for ICA-AtoM, and was originally referred to as “Qubit-OAIS”. Over time, though, the development team recognized that the direct association with ICA-AtoM may be too exclusive, obscuring the larger goal to

¹ ISO 14721:2003, Space data and information transfer systems – Open archival information system – Reference model (2003).

² Kevin Bradley of the National Library of Australia with Junran Lei and Chris Blackall of the Australian Partnership for Sustainable Repositories, published “Towards an Open Source Repository and Preservation System: Recommendations on the Implementation of an Open Source Digital Archival and Preservation System and on Related Software Development” for the UNESCO Memory of the World Sub-Committee on Technology.

³ Ibid., 3.

⁴ Ibid., 8.

⁵ Artefactual Systems, Inc., website, Qubit.

allow Archivemata to integrate with other systems. Therefore, qubit-oais became Archivemata, an open-source digital preservation system designed for standards-based, long-term access to digital materials.

2. City of Vancouver Digital Archives Project

With UNESCO, the City of Vancouver Archives was one of the first institutions to allocate resources for Archivemata development. The objective of their project was to establish a prototype digital archives environment and provide direction on the management framework within the City of Vancouver Archives (CVA) to implement and sustain a digital archives. The CVA is responsible for permanently preserving archival records created by the City of Vancouver and its various boards and agencies. It is also responsible for acquiring the archives of private-sector individuals and organizations within the constraints imposed by its acquisitions mandate. Increasingly, many of the records created by these various bodies exist only in digital form. The CVA recognized their responsibility to ensure that it had adequate policy infrastructure and technical capacity in place to be capable of permanently preserving and providing access to authentic and reliable digital records. To meet this responsibility it partnered with Artefactual Systems, Inc. and launched the Digital Archives project.⁶

The Digital Archives project focused on problems related to preserving municipal digital records created within the City's Electronic Records and Document Management System, called VanDocs, as well as digital records created outside of the VanDocs environment, in particular, records created and/or maintained by individuals and organizations from the private-sector in recordmaking and recordkeeping systems that the Archives had no control over. The diversity of records and recordkeeping systems in this prototype project were ideal towards developing a system that could adapt to a multitude of memory institutions with different mandates and acquisition policies. Digital preservation goals may be similar industry-wide, but different types of digital objects and workflows are unique and plentiful.

3. Digital Preservation – An Overview

In modern-day institutions, daily operations and communications are managed through the creation and exchange of digital information (e.g., business records, email, technical drawings). However, unlike paper records, which can sit untouched in boxes or filing cabinets for years or even decades without harm, digital records require specialized actions to manage and preserve them. In fact, the long-term accessibility, usability and authenticity of digital materials are at risk due to the inherent fragility and complexity of digital objects and to technological incompatibilities or obsolescence at the level of file storage, application software, metadata and file formats.

Digital records can easily be lost, deleted or modified; sometimes maliciously but more often through mishap (e.g., storage media failure, lack of proper backups in case of accidental deletion) or a simple lack of understanding that they are records of the organization that should be handled with as much care and attention as paper records have been in the past. Over time, some formats can no longer be

⁶ Vancouver Digital Archives wiki, last accessed September 2, 2012, http://artefactual.com/wiki/index.php?title=Vancouver_Digital_Archives.

read when the software that created them is upgraded or discarded. This can result in serious productivity issues or lost business opportunities to re-purpose and re-use digital assets.

Even if they can be read, the rendering of digital files may not be reliable; the “look and feel” may be altered or there may be data loss due to the fact that they are being rendered using different software or a newer version of the software that created them. Moreover, the records can easily become detached from their context: that is, they can be separated from their metadata or lose their links to other records that were originally created and maintained as part of the same business process. This means that even if an electronic record can be retrieved and read it may have compromised its authenticity and its evidentiary value in legal and regulatory proceedings.

For these reasons, the Archivematica project focused on maintaining accessibility, usability, and authenticity of digital information objects over space, time, and technology. To accomplish the task, they set about to build their system in compliance with the ISO-OAIS functional model and other digital preservation standards and best practices.

4. OAIS Functional Model Analysis

It was with the aforementioned digital preservation goals in mind that Artefactual Systems and the CVA began building what the UNESCO Sub-Committee had imagined. In late 2008, Artefactual and the CVA project team began conducting a comprehensive requirements analysis to establish minimal baseline functional requirements, policies and procedures for a digital archives system based on accepted standards. The initial round of requirements gathering started with the development of use cases based on the ISO-OAIS model.⁷

The OAIS is the de-facto standard for designing digital archives systems. Many digital preservation systems or projects claim to be “OAIS-compliant” and this was also a goal for the Vancouver Digital Archives project, but at the time it was difficult to trace requirements between the OAIS standard and systems that claim to be “OAIS-compliant”. The detailed OAIS requirements analysis with its use case methodology to establish what the system requirements are for Digital Archives to be “OAIS-compliant” was an attempt to build traceability into the Archivematica project.

A simple use case methodology⁸ was established to structure the use cases. Use cases were clustered around the same broad categories as the OAIS Functional Entities. Like the latter, use cases were organized into hierarchy with high-level scenarios broken down into more specific tests (sub- and sub-sub-cases). The use cases attempted to present plain language descriptions of what a Digital Archives should accomplish.

Functional, metadata and technology requirements were derived from the use cases and from an open-source technology evaluation. The functional requirements specified what Archivematica should be able to do. Metadata requirements stipulated what data attributes had to be captured for each step. Technical requirements stipulated specific technical features, formats or protocols that had to be implemented. Policies and procedures were also derived from the use cases, in that they are developed to support all the steps the use cases contain. For example, a use case may state: “System implements

⁷ City of Vancouver Digital Archives wiki, Requirements Analysis, last accessed September 2, 2012, http://artefactual.com/wiki/index.php?title=Requirements_Analysis.

⁸ Ibid., Use Case Methodology, http://artefactual.com/wiki/index.php?title=Use_Case_Methodology.

disaster recovery policies such as duplication to remote storage facility”; successful completion of this step naturally requires the development of such policies.

The functional requirements were expressed as UML Activity Diagrams. A first set of these was based directly on the OAIS use cases without any additional interpretation. These were then revised as a second set of CVA-specific Activity Diagrams based on a business process and IT architecture analysis carried out by the project team as well as the ongoing technology and tools evaluation⁹ and software integration and development work.¹⁰

In the course of the requirements analysis, the project team had an opportunity to become a part of the InterPARES 3 Project.¹¹ The team consulted with the InterPARES 3 Project to conduct a gap analysis between OAIS and the InterPARES 1 Project’s Chain of Preservation (COP) Model.¹² Review of the model, along with consultations with archivists about processing analogue records, revealed that appraisal occurs in a few different stages during archival processing. This gap analysis led to use cases and UML Activity Diagrams which addresses appraisal requirements for Archivematica.

5. Archivematica

The thorough use case and process analysis by CVA and Artefactual identified workflow requirements to comply with the OAIS functional model. The resulting Archivematica system uses a micro-services design pattern to provide an integrated suite of free and open-source software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. It allows digital preservation professionals to process digital transfers (accessioned, simple and complex digital objects), arrange them into Submission Information Packages (SIPs), apply media-type preservation plans and create high-quality, repository-independent Archival Information Packages (AIPs). Archivematica is designed to upload Dissemination Information Packages (DIPs) containing descriptive metadata and web-ready access copies to any access system (e.g., DSpace, ContentDM, ICA-AtoM, etc.). Users monitor and control the micro-services via a web-based dashboard.

Through deployment experiences and user feedback, including the gap analysis conducted with the InterPARES 3 Project, Archivematica has expanded beyond OAIS to address analysis and arrangement of transfers into SIPs and allow for archival appraisal at multiple decision points. The Archivematica micro-services implement these requirements as granular system tasks which are provided by a combination of Python scripts and one or more of the free, open-source software tools bundled in the Archivematica system.

Archivematica uses METS, PREMIS, Dublin Core and other recognized metadata standards. The media type preservation plans it applies are based on an analysis of the significant characteristics of file formats.¹³ Archivematica supports emulation preservation plans by preserving original bitstreams, and it will support migration preservation plans by monitoring at-risk file formats and providing a process to migrate them at a future date. Nevertheless, Archivematica’s default preservation strategy is to normalize

⁹ Ibid., Technology/Tools Evaluation, http://artefactual.com/wiki/index.php?title=Technology/Tools_Evaluation.

¹⁰ Ibid., Software Integration/Development, http://artefactual.com/wiki/index.php?title=Software_Integration/Development.

¹¹ InterPARES 3 Project, last accessed May 21, 2012, http://www.interpares.org/ip3/ip3_index.cfm.

¹² InterPARES 2 Project, Chain of Preservation (COP) Model, last accessed May 21, 2012, http://www.interpares.org/ip2/ip2_model_display.cfm?model=cop.

¹³ Archivematica wiki, Significant characteristics evaluation, last accessed August 3, 2012, https://www.archivematica.org/wiki/Significant_characteristics.

digital objects into preservation formats upon ingest, in order to make best use of the limited time that organizations will have to process and monitor large, diverse collections of digital objects. Building normalization paths into the software requires choosing target formats and integrating open-source tools to perform the migrations. The choice of preservation formats is based on four basic criteria, which will be familiar to many of those who have experience with digital preservation:

1. The specification must be freely available.
2. There must be no patents or licenses on the format. Archivemata's preservation formats are all open standards.¹⁴
3. Other established digital repositories should be using or have endorsed the format.
4. There should be a variety of writing and rendering tools available for the format.

Selection of preservation formats has been an iterative process of researching best practices, testing normalization tools, and, as far as possible, comparing before and after results of conversions by measuring significant properties. The choice of access formats is based on the ubiquity of viewers for the file format as well as the quality of conversion and compression.

Archivemata prepares a METS file for each SIP and packages it with the AIP. The purpose of the METS file is to capture, in a standardized way, information about all the objects that are being preserved. The METS file lists all of the objects in the AIP, categorizes their role (original, preservation copy, submission documentation, etc.), and allows an original object to be intellectually linked to its preservation copy. The METS file also includes a robust PREMIS (Preservation Metadata Implementation Strategies) implementation which provides highly detailed technical information about each object, an audit trail of actions taken on the object since it was ingested, and detailed and granular rights information.

6. Open-Source Software and Agile Development Methodology

All of the software, documentation and development infrastructure are available free of charge and released under AGPL3 and Creative Commons licenses to give users the freedom to study, adapt and re-distribute these resources as best suits them. Archivemata development is led by Artefactual Systems, a Vancouver based technical service provider that works with archives and libraries to implement its open-source solutions as part of comprehensive digital preservation strategies. All funding for Archivemata development comes from clients that contract Artefactual's team of professional archivists and software developers to assist with installation, integration, training and feature enhancements. The majority of Archivemata users take advantage of its free and open-source license without additional contracting services.

Archivemata follows an agile software development methodology. Its micro-services model is malleable enough to allow for a rapid release cycle and iterative, granular updates to the requirements documentation, software code and end-user documentation. Artefactual clients and the Archivemata user community help to prioritize new features and bug fixes for each release.

¹⁴ Wikipedia definition of open standards, http://en.wikipedia.org/wiki/Open_standard.

7. Pilot Project Prototyping and User Experience at the International Monetary Fund

One of the first Archivemata project clients was the International Monetary Fund (IMF). The IMF joined Artefactual and the City of Vancouver Archives in real-world testing of Archivemata. The IMF's experiences as users, as well as the confidentiality issues of digital archives the IMF is working through, have been invaluable to Archivemata's improvement.

Based in Washington, DC, the IMF is focused on analysing and reporting on world economic conditions, and providing loans to member countries when needed. The IMF Archives provides the institutional memory of the Fund. Their paper collections date back to the Bretton Woods Conference in 1944 which created the Fund and continue up to the present day. However, the IMF's digital archives is much newer. Though the Fund's network drives contain documents dating back to 1980, and external media up to and including boxes of punch cards have been found while processing paper collections, it's only been in the last few years that the Archives has been able, from a funding and expertise perspective, to begin to work on how to bring those records into the fold, preserve them, and make them accessible.

As Digital Archivist, Paul Jordan took the lead in the hands-on gathering and testing of collections. One of the main tools he used in his investigations was Archivemata, which was first installed in December, 2009 and has been updated regularly ever since. IMF Archives used it in both prototype and pilot projects with a variety of source systems and file types, including legacy shared drives, current email mailboxes, and external media. All of the documents used came from actual collections brought in from various Fund departments.

Installation was simple since Artefactual uploaded the entire Archivemata platform to the Ubuntu repository, and so theoretically all you need is an Ubuntu machine and an internet connection, and a few commands will download the entire thing. In fact, the CVA got a couple of older computers which were on their way to the recycling center after a computer refresh, sat them on a table, installed Ubuntu, and then downloaded Archivemata. After adding Ethernet cables to link them together, they had a digital processing cluster on a table. Unfortunately, the IMF does not work that way. The IMF is a very security paranoid organization, and with good reason. No software is allowed into our IT environment, not even the development environment, without first putting it through a security accreditation process. This provided a challenge for the IMF Archives and Artefactual, since the software had not matured enough, and the project schedule was too short, to accommodate full testing.

Ultimately, IMF Archives created an isolated sandbox, a couple of virtual computers on the IMF's virtual server farm that were completely separated from the rest of the IMF's network, with the sole exception of one link out to a single network share which could be used to load files into and out of the sandbox. This worked fine from a security perspective, but it makes installation a bit cumbersome. One of the things that is very specifically blocked was any kind of internet access, which meant staff could not simply download Archivemata from the repository. Archivemata and all of its dependencies had to be loaded onto external media and then moved into the sandbox. However, during the most recent installation, IT was able to temporarily open a port to the internet, and then shut it down as soon as the installation was done. That installation took about half an hour.

Over the course of our pilot projects, the IMF Archives focused on three source systems: departmental network drives, external media, and email. The network drives allowed for the widest variety of content, with files dating back to 1980, many of them in formats that were difficult to identify even with tools like Jhove and Droid. Files from the external media were much the same, though complicated by balky and sometimes corrupt media. The emails came from a departmental shared mailbox and were the

most sensitive, as well as being the largest files. There were hundreds of files from the shared drives, but individually few of them were more than a couple hundred kilobytes, and gigabytes of email.

From its earliest days, Archivematica has been an invaluable tool for donor out-reach. Even when it was still crude, it provided the IMF Archives with a concrete reason to reach out to departments in the Fund with collections of interest. The IMF Archives is a corporate archive without any external donors. All of their collections come from departments within the Fund. While this means that many departments already know the Archives exists, donor relations and networking is no less critical. Policies, procedures, and contacts for paper records are well established. Those for digital have barely begun, and while the existing Archival contacts form a firm groundwork, the specifics of transfer are still to be decided.

But once the IMF Archives had a pilot project with an actual working system that needed sample collections, it gave the Archives the perfect opportunity to talk to departments of interest. Outreach was overall successful. Some potential donors were extremely interested, and the Archives made friends in more than one department. People are starting to become aware of digital preservation issues, or at least worry about losing access to files and email, and they seem very willing to work with the Archives.

One hazard, however, is promising too much, too soon. The Archivematica version installed when these interviews occurred was not yet production ready. The Archives were not ready, either. It was a pilot project, and it was important to stress that to the people we were working with. It made for a very interesting tightrope to walk: trying to convince them that the Archives knew what they were doing, and that their records would be safe in their hands, without going so far as to take formal custody of the items and promise full archival processing and then access; things they did not yet have the procedures, the software, or the personnel to deliver.

For small and medium-sized institutions, Archivematica is a platform that ties together many smaller services, each with its own set of tools. When digital archivist Paul Joran got into the field of Digital Archives, he spent an entire summer trying to install and get working a few of the individual tools contained within Archivematica. However, he had extremely limited IT assistance and was not able to make much headway. In contrast, the IMF was able to acquire a single package that contained all of the software he had struggled with and more, all working together in the same direction already. IMF Archives was able to test the OAIS model against what they had and what they wanted to do with version 0.5 of Archivematica.

Archivematica is also a very flexible system that can support whatever workflows an organization might have. For the IMF Archives, one of those workflows that requires a great deal of effort and that is only just beginning to be addressed in the digital archives is classification.

The IMF is a very security-conscious organization. The Archives has a full-time declassification archivist whose role is the identification, removal, and processing of classified materials within the paper archive, and none of the collections can be made available until this screening has been completed. The same thing will be true for digital objects.

Up to this point, the Archives has focused on records already open to the public. This has primarily been digitized archival collections: the entire repository of scanned documents from the Executive Board that are open to the public, some of the archival country files, and an oral history collection. Everything that has been digitized has already been screened and declared open; the Archives does not scan anything still confidential. No born digital collections have been made available yet, because it is in those extremely large legacy collections that will likely cause problems. When IMF Archives first started using Archivematica, it was not set up to handle both public and non-public documents, because Artefactual had never worked with a partner that had addressed requirements for such a need. A lot of the features

around classification and document review were suggested by the IMF. Fortunately, the sandbox approach allowed the Archives to work with classified documents in isolation and to do some analysis.

Some of the documents at the IMF already have security classifications assigned to them. The Fund's document management system tracks classification, and there is also an Outlook add-in that does the same for email. Those will be the easy ones to identify. It will also be easy to determine which of the documents within that time period are public, or subject to automatic declassification, because lower classification statuses are also tracked. Older records from before these systems were implemented are more of a problem. Email will usually have classification in the header, but documents, for instance off of shared drives, will generally have to be opened to determine whether or not they're classified. Also, many emails have attachments that are themselves classified, and the classification of the two does not always match.

Therefore, one of the Archive's steps during appraisal will be a macro review of classification status. The hope is that based on provenance, archivists will be able to get a general idea of the quantity of classified documents, which can then factor into a collection's processing priority. It can also help divide SIPs for ingest; if a processing archivist can determine that everything outside a particular sub-directory is open, the classified subdirectory can be sequestered and everything else made available. Once the macro declassification appraisal has been completed, the collection will be appraised and processed normally. Only once processing is complete will archivists go back and do a second, item-level classification screening.

One of the things that may ease the process, and Archivemata is planning to implement for their 1.0 release, is full-text indexing of incoming documents. This will allow archivists to search for classification keywords and phrases. There is a large, though finite, number of terms that can classify documents; if staff can identify those documents that contain those words, they will be able to weed out a significant number of open documents. However, they will still need someone to go through the documents that have been flagged and determine whether each really is a "secret document," or whether it's an email where the sender is talking about his kids wanting a "secret moon base."

8. Archivemata 0.9 Beta Release Features and 1.0 Development Roadmap

Beyond the IMF and CVA, Artefactual clients include the University of British Columbia Library, Simon Fraser University Archives, and the Rockefeller Archives Center. Based on their input, Artefactual's own research and goals, evolving best practices and requirements for digital preservation systems and the input from our user community at-large, our recent release included numerous improvements on the previous iterations. Working with pilot project implementers, the Archivemata team has gathered requirements for managing a backlog of indexed digital acquisitions transfers, creating a SIP from a transfer or set of transfers, basic arrangement and description, preserving email, and receiving updates about new normalization paths via a format policy registry (FPR). After creating workflows that would account for real-world archival processing needs, these requirements were added to our development roadmaps for 0.9, 1.0 and subsequent Archivemata releases.

The first Archivemata beta release, Archivemata 0.9, became available for download from the Archivemata website in early September of 2012. In addition to fixing bugs and enhancing features, release 0.9 includes the following new features:

- An update to the Ubuntu 12.04 LTS as the base operating system.
- The web browser dashboard interface has replaced most of the file browser functionality.
- DIPs can be uploaded to CONTENTdm.

- All AIP metadata can be indexed and searched using a tool called ElasticSearch.
- The rights module is updated to the most current PREMIS implementation, PREMIS 2.2.
- Email handling is improved and there is a prototype ingest of maildir.
- User accounts can be created.
- Automatic restructuring of transfers for compliance.
- In the dashboard, jobs are grouped into micro-services.
- Ingest of Library of Congress Bagit format.
- Nightly backup of MCP MySQL database.
- Scalability enhancements.

Users and clients will continue testing and processing digital records using the 0.9 release, all the while informing the 1.0 release in early 2013. So far, there is a base set of features and enhancements for that release on the Archivematica wiki. Proposed features for 1.0 so far include:

- Develop a Format Policy Registry (FPR) and upload/download of format policy information between FPR and Archivematica instances.
- Upgrade file identification used as the basis to trigger format policy actions (aka ‘preservation plans’).
- Include a manual normalization workflow.
- Improve email handling.
- Add ability to edit format policies from preservation tab in the dashboard.
- Add ability to add/change format policies from FPR updates.
- Add a workflow for applying updated format policies to pre-existing AIPs.
- Include advanced search screens for searching AIP contents in the dashboard.
- Generate DIPs from the access tab in the dashboard.
- Include visualization of transfers.
- Include file-level Dublin Core and rights metadata entry.
- Include field validation in rights templates.
- Index transfers and identify/flag personal information. Evaluate BitCurator tool to determine how much functionality/data can be integrated/re-used prior to Archivematica ingest.
- Customize statistical reporting
- Improve AIP retrieval (whole or part) and delivery.
- Possibly remove packaging/compression for AIPs .
- Sync metadata between DIP and AIP for CONTENTdm, AtoM via OAI-PMH.
- Include AIP versioning (METS file updates).
- Include an enterprise ID service to connect AIPs and DIPs (dns/uuid).
- Enhance CONTENTdm DIP upload.
- Allow DIP upload to XTF.
- Transfer metadata from Archivist Toolkit.
- Ingest of TRIM exports.
- Automatic ingest from DSpace using OAI-PMH - OAI API for Archivematica dashboard.
- Management of persistent MCP metadata that does not end up in AIP.
- Make MCP processing workflows editable from the administration tab.
- Improve multi-node processing.
- Further scalability testing/prototyping.

9. Next Steps

The Archivemata project analysis and development described in this article are driven by practical demands from our early adopter community, including the initial vision from the UNESCO Memory of the World Sub-Committee on Technology. The alpha release prototype testing sponsored by our contract clients and shared by a growing community of interested users from the archives and library professions and beyond has provided the opportunity to spearhead the ongoing evolution of digital preservation knowledge in the form of a software application that is filling a practical need for digital preservation professionals.

At the same time, the digital curation community is also evolving and maturing. New tools, concepts and approaches continue to emerge. The Archivemata technical architecture and project management philosophy are designed to take advantage of these advancements for the benefit of Archivemata users and the digital curation community at large.

The free and open-source, community-driven model provides the best avenue for institutions to pool their technology budgets and to attract external funding to continue to develop core application features as requirements evolve. This means the community pays only once to have features developed, either by in-house technical staff or by third-party contractors such as Artefactual Systems. The resulting analysis work and new software functionality can then be offered at no cost in perpetuity to the rest of the user community at-large in subsequent releases of the software. This stands in contrast to a development model driven by a commercial vendor, where institutions share their own expertise to painstakingly co-develop digital preservation technology but then cannot share that technology with their colleagues or professional communities because of expensive and restrictive software licenses imposed by the vendor.

Roles and Responsibilities in Digital Preservation Decision Making

Towards Effective Governance

Hannes Kulovits,¹ Christoph Becker² and Andreas Rauber²

¹Austrian State Archives; ²Vienna University of Technology

Abstract

In this article, we take a critical look at the current state of the art in decision making for digital preservation operations. The goal of preservation planning is to ensure that the optimal decision is taken to maintain the authenticity and understandability of digital objects. To accomplish this, the preservation planner needs to have an understanding of both the organizational context and the challenges posed by the quest for digital longevity. Clear roles and responsibilities for each process are a key success factor of effective governance. Hence, we elaborate on required activities and discuss roles and responsibilities. The conclusions shall contribute to a clarification of the planner role and highlight crucial skills and expertise required.

Authors

Hannes Kulovits is head of the Digital Archive division at the Austrian State Archives, where he is responsible for operations, further development, compliance and certification of the digital long-term repository. The long-term preservation of electronic government records is one of his main concerns, which also leads him to an active involvement in preservation research. His research focus lies on automation of preservation planning activities and policies. He has been involved in a number of projects co-funded by the European Commission including DELOS, DPE, PLANETS, and TIMBUS.

Dr. Christoph Becker is Senior Researcher at the Vienna University of Technology. His interests focus on the areas of Information Systems, Digital Preservation, Information Management, Decision Analysis, and IT Governance. He has published extensively on trustworthy decision making and control in Digital Preservation. As architect of the preservation planning method and tool Plato, he was shortlisted for the DPC's DP Award 2010. He is member of ACM, IEEE and ASIST and has been involved in the European research projects DELOS, PLANETS, DPE, and SHAMAN. Currently he is leading the sub-project Scalable Planning and Watch of the FP7-funded project SCAPE.

Dr. Andreas Rauber is Associate Professor at the Department of Software Technology and Interactive Systems (ifs) at the Vienna University of Technology (TU-Wien). He is president of AARIT, the Austrian Association for Research in IT and a Honorary Research Fellow in the Department of Humanities Advanced Technology and Information Institute (HATII), University of Glasgow. His research interests cover the broad scope of digital libraries and information spaces, including specifically text and music information retrieval and organization, information visualization, as well as data analysis, neural computation and digital preservation.

1. Introduction

Decisions are required on a variety of levels in any organization concerned with a long-term view on the value of digital information, ranging from decisions about long-term strategies and the scope of preservation to the tactical level of preservation operations. This article takes a critical look at the current state of the art in decision making and governance processes for operational digital preservation. At the core of digital preservation is the question of information preservation. It focuses on the search for the

optimal way to achieve longevity of information for a certain target group. The constant changes in technologies, user communities, and organizational context require active involvement to achieve this. In this light, a *preservation action* is a concrete action, usually implemented by a software tool, performed on specific content in order to achieve preservation goals. For instance, a preservation action can consist of the migration of content to a different format using a certain tool in a certain configuration and software and hardware environment. Operational preservation thus searches for the optimal preservation action that ensures the authenticity and understandability of specific digital content for certain users.

Frequently, preservation decisions are now taken by (small) teams of people specifically tasked with digital preservation, which is considered as a highly specialized, focused competence. However, these decisions concern very different aspects of preservation, ranging from high-level considerations about regulatory compliance and business strategies to low-level IT concerns about Quality Assurance of metadata transformations.

Governance “refers to the way the organization goes about ensuring that strategies are set, monitored, and achieved.”¹ As such, governance sets the institutional and policy framework in an organization. Governance frameworks in Information Systems show that understanding the roles and responsibilities for each process is a key success factor of effective governance.² In this light, we outline typical chains of decisions in a preservation environment and illustrate corresponding tasks and roles in the primary dimensions of business/technology versus strategy/operations. It becomes clear that a transparent and explicit assignment of roles and responsibilities as well as a definition of expected skills and expertise for these tasks and activities is required. This applies in particular to the *preservation planner*. Given the current state of the domain of digital preservation, it is not surprising that a full understanding of the planning role has yet to be formed. However, it is clear that a successful preservation planner needs to have an understanding of the business context and goals and acquire in-depth knowledge of the technical intricacies to be resolved.

The article is structured as follows. We will shortly outline the context of digital preservation and preservation planning and clarify the scope of planning. We review experiences in a preservation planning case study in the light of the stakeholders involved, and discuss the various facets that arise in a standard preservation planning activity. We show that preservation planning needs to be positioned on an operational level, with clear goals, constraints and responsibility assignments as a prerequisite to success. We outline the tasks and activities that form part of the process, discuss the expertise and skills required for each of these, and reflect on the role of the preservation planner. We further draw conclusions about gaps that should be addressed and modeled more explicitly to support organizations in specifying their digital preservation governance processes. These conclusions shall contribute to a clarification of the state of the art and practice in digital preservation decisions and support prospective adopters of systematic preservation planning in analysing their readiness for transparent governance processes.

¹ Kenneth G. Rau, “Effective Governance of IT: Design objectives, roles, and relationships,” *Information Systems Management* 21, no. 4 (2004): 35-42.

² IT Governance Institute, “COBIT 5 – A business Framework for the Governance and Management of Enterprise IT,” 2012.

2. Digital preservation and preservation planning

Management of the diverse risks to the longevity of digital information requires awareness and treatment of threats and vulnerabilities on three different levels, the physical, logical and semantic level. On the physical level, the raw bit streams have to be preserved over time. Additionally, a correct way to interpret this bit-stream must be preserved as well, which arguably is the more challenging part of digital preservation: Digital objects require specific program versions to open and render them; these in turn depend on specific software components and an operating system, which in turn runs on and supports a specific type of hardware components. A consumer of content in turn will access objects using a specific environment that needs to support the object at hand. If any of the elements in the performance chain is lost, a digital object cannot be rendered successfully and is reduced to a useless concatenation of zeros and ones. It becomes clear that even having a storage medium being capable of retaining digital data for a millennium is worthless if the means of interpretations are lost. A comprehensive overview of the challenges in digital preservation and of preservation strategies is provided in the accompanying document to the UNESCO charter for the preservation of digital heritage.³

Digital curation as “[t]he active involvement of information professionals in the management, including the preservation, of digital data for future use,”⁴ covers the entire lifecycle of a digital object from the early stages of conceiving and planning it to either its disposal or long-term preservation and possible re-use. Preservation is thus one in the entire set of curation activities.⁵

The Reference Model for an Open Archival Information System (OAIS)⁶ describes, on a high level, the functional and information model of an archival information system and the exchange of information between the different entities. Furthermore it lists roles involved and their responsibilities; the three main roles described in the OAIS are *Producer*, *Consumer*, and *Management*. While Producers constitute “persons or client systems that provide the information to be preserved,”⁷ the Consumers’ main concern as “persons, or client systems who interact with OAIS services” is to “find preserved information of interest and to access that information in detail.”⁸ Management is “[t]he role played by those who set overall OAIS policy as one component in a broader policy domain.”⁹

A core function in the model is Preservation Planning, which

...provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored

³ Colin Webb, “Guidelines for the Preservation of Digital Heritage,” prepared by the National Library of Australia for the Information Society Division, UNESCO, March 2003, <http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>.

⁴ Elizabeth Yakel, “Digital curation,” *OCLC Systems & Services: International digital library perspectives* 23, no. 4 (2007): 335-340.

⁵ Philip Lord and Alison Macdonald, “e-Science Curation Report: Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision,” prepared for The JISC Committee for the Support of Research (JCSR), 2003, http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf.

⁶ International Standards Organization, *ISO 14721:2012—Reference model for an open archival information system (OAIS)*, The Consultative Committee for Space Data Systems, Recommended Practice, CCSDS 650.0-M-2, June 2012, <http://public.ccsds.org/publications/archive/650x0m2.pdf>.

⁷ Ibid., p. 1-14.

⁸ Ibid., p. 1-10.

⁹ Ibid., p. 1-13.

in the OAIS remains accessible to, and understandable by, the Designated Community over the long term, even if the original computing environment becomes obsolete.¹⁰

A preservation plan then

...defines a series of preservation actions to be taken by a responsible institution due to an identified risk for a given set of digital objects or records (called collection). The Preservation Plan takes into account the preservation policies, legal obligations, organizational and technical constraints, user requirements and preservation goals and describes the preservation context, the evaluated preservation strategies and the resulting decision for one strategy, including the reasoning for the decision. It also specifies a series of steps or actions (called preservation action plan) along with responsibilities and rules and conditions for execution on the collection. Provided that the actions and their deployment as well as the technical environment allow it, this action plan is an executable workflow definition.¹¹

This preservation planning can be supported by tools such as the planning tool Plato,¹² which implements the planning method described in Becker et al. 2009.¹³ The publicly available tool guides decision makers via a structured workflow to create an actionable preservation plan for a well-defined set of objects which are considered being at risk, based on a thorough goal-oriented and evidence-based evaluation of potential actions. The workflow comprises the following phases:

1. *Define requirements:* In the first phase, goals and criteria are specified that the optimal preservation action needs to fulfill. The specification starts with high-level goals and breaks them down into quantifiable criteria, thus creating an objective tree. The objective tree forms the basis for evaluating the preservation actions.
2. *Evaluate alternatives:* Empirical evidence for evaluation of all potential candidate solutions is gathered via controlled experimentation. All alternative candidates are applied to real sample content selected from the set of objects to be preserved and evaluated according to the specified set of criteria (i.e., for every criterion, a measure is collected for each experiment).
3. *Analyse results:* To allow comparison across different criteria and their measurements, a utility function is defined for each criterion. This utility function maps all measures onto a uniform utility scale. Relative importance factors on each level of the goal hierarchy model the preferences of the stakeholders. An in-depth analysis of the resulting performance of candidates (i.e., their weighted utilities throughout the goal hierarchy) leads to an informed recommendation of an alternative.
4. *Build preservation plan:* In this phase, concrete steps required to put the action plan into operation are defined. This not only includes an accurate and understandable description of on which preservation action is to be executed on which digital objects the and how, but also the

¹⁰ Ibid., p. 4-2.

¹¹ Christoph Becker, Hannes Kulovits, Mark Guttentbrunner, Stephan Strodl, and Andreas Rauber, "Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans," *International Journal of Digital Libraries* 1, no. 2 (2007): 92-101, <http://www.ijdc.net/index.php/ijdc/article/view/27/16>.

¹² <http://www.ifs.tuwien.ac.at/dp/plato>

¹³ Becker et al., "Systematic planning for digital preservation."

quality assurance measures to be taken along with it to ensure that the results are corresponding to expectations. Furthermore, procedures and responsibilities for plan execution are defined.

As opposed to strategic planning, which is of a conceptual nature and inherently focused on a vision, the result of preservation planning is an *operational* plan with a concrete, focused preservation action fulfilling clear objectives. These objectives are focused on the overarching goal of ensuring authentic access to specific content in understandable form for a specified user group. While an organization will typically define one strategic plan, it will very often hold more than one homogeneous set of objects of interest for more than one group of users. This means that a set of preservation plans will be required to specify concrete actions to take for keeping the sets of digital objects alive over time according to the organization's strategy and policies, and these plans will evolve according to different lifecycles than the strategic plans. This distinction is crucial to ensure proper alignment of preservation planning to the strategies and policies of the organization.

Preservation policies in turn have been discussed on different levels. Criteria such as the ISO 16363 Repository Audit and Certification Criteria¹⁴ aim for verifying the compliance of an archive to what are perceived as standard "best practices". On an operational level, executable rules such as those described in MacKenzie and Reagan, 2007¹⁵ aim at operational control of preservation systems and support monitoring the compliance of a system to specified constraints. For decision making, however, policies are non-enforceable elements of governance that guide, shape and control the strategies and tactics of an organization.¹⁶ This is also the perspective we adopt in this article where we speak about policies.

3. Preservation planning in practice

The Bavarian State Library (BSB) has been amongst the first institutions to actively deploy the planning process discussed above in an organization. Triggered by the observation that institutions such as the British Library decided to move forward towards migrating parts of their collections to JPEG2000, the BSB questioned the current file format of choice for high-quality scans, which constitute one of their largest digital collections.

Staff from the library embarked on a quest to find the optimal file format for this particular type of content and prepare a preservation plan.¹⁷ Storing image files without using compression makes them more robust against bit corruption. However, going without compression has to be balanced against incurring storage costs. The state library thus commenced planning with the specific goal to evaluate the option of migrating to JPEG2000, i.e., evaluating whether the BSB would benefit from migrating their TIFF collections to JPEG2000.

¹⁴ International Standards Organization, *ISO 16363:2012—Space data and information transfer systems -- Audit and certification of trustworthy digital repositories*, The Consultative Committee for Space Data Systems, Recommended Practice, CCSDS 652.0-M-1, September 2011, <http://public.ccsds.org/publications/archive/652x0m1.pdf>.

¹⁵ MacKenzie Smith and Reagan W. Moore, "Digital Archive Policies and Trusted Digital Repositories," *International Journal of Digital Curation* 1, no. 2 (2007): 92-101, <http://www.ijdc.net/index.php/ijdc/article/view/27/16>.

¹⁶ Object Management Group, "Business Motivation Model 1.1," 2010.

¹⁷ Hannes Kulovits, Andreas Rauber, Markus Brantl, Tobias Beinert, and Anna Kugler, "From TIFF to JPEG2000? Preservation planning at the Bavarian State Library using a collection of digitized 16th century printings," *D-Lib Magazine* 15, no. 11/12 (November/December 2009), <http://dlib.org/dlib/november09/kulovits/11kulovits.html>.

A series of different people were involved in the preservation planning process. While two people from the library staff were in charge of proceeding through the steps of the planning workflow in general, the process of gathering the preservation goals and criteria involved a number of people from within the organization and outside. The people involved in key planning tasks were the following.

- The head of the digital library and digitization services was responsible for defining of the planning scope, specifying and clarifying goals and constraints, and approving the preservation plan.
- The person responsible for the digitization process and its implementation contributed knowledge on peculiarities of the digital image files, including metadata of the scanning process.
- The person responsible for storage at the supercomputing center that provides the technical storage facilities for the state library made sure that limitations of the technical infrastructure were considered. This included restrictions on possible preservation action tools and storage limitations. Furthermore, the retrieval and re-ingest process had to be considered with respect to costs and feasibility.
- Two library researchers and a historian were responsible for the identification of significant properties, a comprehensive definition of the user community, and the evaluation of the considered preservation actions.
- Finally, an external preservation planning expert moderated the workshops and guided the decision makers through the steps of preservation planning.

In the beginning, a clear **definition of the planning scenario** had to be created. This definition specifies a certain set of digital objects and the user community for whom its accessibility is of concern. In this case, the focus was on high-quality scans of 15th and 16th century incunabula, which are made accessible in a low-resolution copy to the general public via the internet. Reproductions of the original are produced using the high-resolution master file.

Once the scenario has been defined, the **context** in which the preservation plan operates needs to be documented; legal obligations/restraints, organizational workflows, and policies relevant to this plan need to be documented. In this case, certain policies of the agency funding the digitization of the incunabula had to be considered. For instance, requirements for the quality of the digital copy had to be respected.

To understand the risks facing the content and describe the scenario at hand, the organization must create a **content profile** describing technical characteristics such as file formats, format versions, date of creation, and the number of embedded objects ('Know what you have'¹⁸). In this case, the collection encompasses more than four million pages of high-quality scans, which were digitized in the course of a funded project between 2007 and 2009. All master files are stored as TIFF version 6 without compression to enable reproductions as close to the original as possible. The collection measures 72 Terabyte.

The search for the optimal action to take continues with specifying the **goals and objectives** that should be met. These need to be collected from a variety of stakeholders and have to be specified in a quantifiable way, starting at high-level objectives and breaking them down into measurable criteria (e.g., bits per sample, Euros per year, frames per second). The resulting objective tree forms the basis of the

¹⁸ Thomas Bähr, Michelle Lindlar, and Sven Vlaeminck, "Puzzling over digital preservation – Identifying traditional and new skills needed for digital preservation," in *World Library and Information Congress: 77th IFLA General Conference and Assembly, San Juan, Puerto Rico, 13-18 August 2011*, <http://conference.ifla.org/past/ifla77/217-bahr-en.pdf>.

evaluation of potential preservation actions. In this case, two half-day requirements sessions were held at the Bavarian State Library to gather the different stakeholder's objectives.

A **set of preservation actions** potentially fulfilling the requirements, need to be selected and the experiment setting determined ('Know what comes ahead'¹⁹). In this case, image conversion tools such as ImageMagick and GraphicsMagick were incorporated into the evaluation.

Carrying out the **experiments** means that each preservation action needs to be applied to each sample object according to the experiment specification. In this case, sample files taken from the collection were migrated using the preservation action tools including ImageMagick. The result was then analysed using characterization tools such as JHove. Using the empirical **evidence** from conducting the experiments, the criteria on the leaf level of the objective tree were then evaluated; the criterion *image size unchanged* for instance could then be evaluated depending on the outcome to either *yes*, or *no*. In order to make the **evaluation values** comparable amongst each other, each criterion in the objective tree was then transformed to a uniform scale between 0 and 5, with 0 being unacceptable and 5 being the best possible evaluation. An essential step is taken here: **Acceptance criteria** are defined and clearly state the constraints the institution is willing to accept. Aggregation of these values over the tree hierarchy leads to a directly **comparable performance value** at root level for each preservation action, with a higher performance value indicating better overall performance. The interested reader is referred to Becker et al., 2009²⁰ for detailed information on the aggregation methods. The entire evidence aggregated to a comparable performance value for each alternative action enables a well-documented and **informed decision** for the preservation action scoring highest performance value.

Evaluation of the potential preservation actions against these objectives resulted in the recommendation to keep the files in their original configuration (TIFF 6, without compression). While this decision kept the status quo, it was the result of an informed and accountable decision-making process specified in a **standardized preservation plan**. The benefits of conversion were at this point outweighed by the costs and risks. The decision was scheduled for review at a later point in time to make sure that potential changes in decision factors will be considered.

All the activities described above had to be carried out by the respective responsible person and documented accordingly. Since the creation of this preservation plan was the organization's first structured approach to finding the optimal solution for a preservation problem, top management had to be called in for certain decisions in some cases. In particular, the organization's policies and strategies had not been fully formed yet at that time. Hence, this first planning activity also laid the ground for subsequent planning efforts for other collections.

The key questions that arise during planning are summarized in Table 1, together with the problem areas they touch upon. The person responsible for planning needs to have an understanding of these areas to be capable of leading the process.

The first phase focuses on a deep understanding of the current situation, i.e., the organizational context, the content and its properties, formats and associated risks, as well as the goals and objectives to be achieved. This is in many ways the crucial phase for planning, and requires the decision makers to understand how they can make their high-level goals and objectives operational to enable informed decisions. For instance, for the Austrian State Archive, the preservation of pre-written official speeches created by the Federal President or his/her employees are of particular historical interest. Such documents

¹⁹ Ibid.

²⁰ Becker et al., "Systematic planning for digital preservation."

Table 1. Key questions and problem areas touched.

Phase	Key questions	Problem areas touched
1: Define requirements	<ul style="list-style-type: none"> • For which digital objects do we create a preservation plan, and why? • Which samples of the objects are representative of the set? • Which are the significant properties of these objects? Who will want to use them, and what are their access requirements? 	<ul style="list-style-type: none"> • Risks to the longevity of digital information • Institutional and Organizational Contexts • User communities • Content profiling and automated analysis • Authenticity • Significant properties • Requirements analysis
2: Evaluate alternatives	<ul style="list-style-type: none"> • Which preservation actions could we apply to keep this content alive and understandable? • What are the effects of applying a certain preservation action? • How can we evaluate software components? • How can we ensure trustworthy decisions? 	<ul style="list-style-type: none"> • Preservation actions • Controlled experimentation • Information sources, evidence, and trustworthiness • Software engineering • Significant properties
3: Analyse results	<ul style="list-style-type: none"> • What are our preferences? • Which are the critical requirements? • Which loss can we accept? • Which costs can we accept? • Which risks can we accept? • Can we achieve our intended goals with the available means within the constraints of our organization? 	<ul style="list-style-type: none"> • Organizational preferences, goals and risks • Multi-criteria decision making • Sensitivity of decision criteria • Authenticity and acceptable loss
4: Define plan	<ul style="list-style-type: none"> • What are the essential steps required to execute the plan as intended? • How can we ensure successful execution of the plan corresponding to specifications? • Who should be responsible for executing the preservation? • Who should be responsible for quality assurance? How much quality assurance is required? • To which degree can we automate preservation actions and quality assurance? • How long should the plan be valid? When do we have to review it? • Which key factors do we need to monitor from now on to ensure we react to critical changes? 	<ul style="list-style-type: none"> • Functional correctness of software tools used for quality assurance • Roles and responsibilities • Service quality • IT operations • Continuous monitoring

often tell a story themselves; they have gone through many iterations until the final version, and are provided with annotations concerning the exact flow of the speech. These documents are commonly stored in the version of Microsoft Word format that was prevalent at the time of creation. Incompatibilities between text editors and their versions make a migration of these documents necessary. This, however, needs to be accurately planned to make sure that the requirements of the different users of these documents are addressed. A high-level goal of the responsible archivist for instance states that

“Documents created by the Federal President or his/her staff about official speeches need to be preserved as they are (including edit history, comments, and remarks regarding intonation).” This requires further breakdown into measurable criteria such as: (1) *Change history must be retained.* (2) *Comments including full name of person and date/time of creation must be preserved.* (3) *Text colour must be retained.*

The second phase requires more technical expertise in conducting experiments that evaluate the feasibility and quality of potential alternatives. In the third phase, on the other hand, the focus lies on organizational preferences, assessment of costs, benefits and risks, and acceptable loss. Finally, the fourth phase requires specification of technical processes and quality management, as well as an understanding of roles and responsibilities in the organization.

Finally, preservation planning is not an isolated procedure. It needs to be accompanied by continuous monitoring of all factors (internal as well as external to the organization) that influenced the planning result. Once a preservation plan has been created and put into operation, quality of service, shifts in designated user communities and their requirements, and the technology environment need to be monitored. Changes should result either in the revision of an existing plan or the creation of a new plan.

4. Preservation planning in context

Preparation of a plan often touches several departments in the organization and thus needs cross-departmental coordination and communication. The desired outcome is the mitigation of an identified risk for a given set of digital objects. The plan provides clear information concerning resource requirements and staffing, quality standards, deployment of the action must comply with, and the schedule of the implementation of the plan.

Many of the decision steps in planning require contextual information. For example, the desirable properties of formats risks and features are generally uniform across an organization, and in fact very similar across organizations. These preferences and the organization’s corresponding risk aversion thresholds need to be supplied to the planning procedure. Such information may be documented in policies addressing external constraints as well as in policies addressing internal goals and directives that control decisions and operations.

This observation highlights the necessity to embed operations and planning in a strong and well-understood organizational context, and align planning to strategic objectives and policies. The perspective hereby needs to be based on a socio-technical system view of a preservation environment, addressing the question “Which capabilities does an organization need for successful preservation”? A capability hereby is an *“ability that an organization, person, or system possesses. Capabilities are typically expressed in general and high-level terms and typically require a combination of organization, people, processes, and technology.”*²¹

The SHAMAN reference architecture for Digital Preservation²² correspondingly provides a contextualized capability-based view on digital preservation, with a strong foundation in Information Systems and Enterprise Architecture frameworks, and thus a more holistic and socio-technical view on the field of digital preservation. It describes goals, drivers and constraints, typical key stakeholders and their concerns, and key capabilities which an organization needs to possess to fulfill its digital preservation mandate. It thus can serve as a guide to understanding governance processes, roles and

²¹ http://pubs.opengroup.org/architecture/togaf9-doc/arch/chap03.html#tag_03_26 (accessed 31 August 2012).

²² SHAMAN Reference Architecture v3.0, “Project Deliverable,” 2012.

responsibilities. Figure 1 shows the relationship between the top-level capabilities, which are grouped in Governance, Business, and Support capabilities. Detailed discussions can be found in SHAMAN, 2012²³ and Becker et al., 2011.²⁴

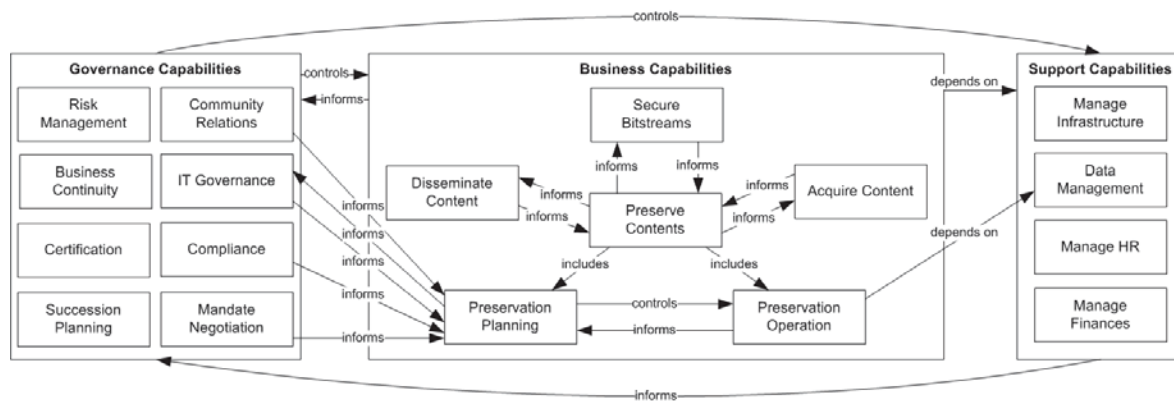


Figure 1. Relationships of digital preservation capabilities.²⁵

Not surprisingly, at the heart of the capability model is the capability **Preserve Contents**, which is the “ability to maintain content authentic and understandable to the defined user community over time and assure its provenance.”²⁶ It is composed of Preservation Operation and Preservation Planning.

Preservation Operation is the “ability to control the deployment and execution of preservation plans. This includes analysing content, executing preservation actions and ensuring adequate levels of provenance, handling preservation metadata, conducting quality assurance, and providing reports and statistics, all according to preservation plans.”²⁷

Preservation Planning is the “ability to monitor, steer and control the preservation operation of content so that the goals of accessibility, authenticity, usability and understandability are met while minimizing operational costs and maximizing (expected) content value. This includes managing obsolescence threats at the logical level as the core risk affecting content’s authenticity, usability and understandability.”²⁸ As emphasized above, preservation planning requires contextual information about the preservation drivers, goals and constraints. This information is managed by a set of **governance capabilities**. As a whole, the governance capabilities manage the scope, context and compliance of the system in order to ensure the fulfillment of the mandate and sustainable operation of the system.

Each of these capabilities must be realized by a combination of organization, people, processes, and technology. For example, the planning tool Plato provides systematic tool support for planning. However, the question remains how the corresponding capability can be specified, created, and assessed in a real

²³ Ibid.

²⁴ Christoph Becker, Gonçalo Antunes, José Barateiro, and Ricardo Vieira, “A Capability Model for Digital Preservation: Analyzing Concerns, Drivers, Constraints, Capabilities and Maturities,” in *Proceedings of the 8th International Conference on the Preservation of Digital Objects (iPRES2011), Singapore, November 1-4, 2011* (Singapore: National Library Board and Nanyang Technological University, 2011), 1-10.

²⁵ Ibid.

²⁶ Ibid.

²⁷ Ibid.

²⁸ Ibid.

organization. This realization of the planning capability requires the development of an organizational process that assigns responsibilities to roles and models the activities and tasks that have to be performed. To better understand the roles and responsibilities that have to be considered, the next section elaborates on the involved stakeholders and roles and discusses the relationship of these roles in a real-world organizational scenario.

5. Stakeholders and roles within the preservation context

In any organization running a digital repository to long-term preserve digital information, a number of stakeholders will be involved in preservation activities. This section thus discusses typical stakeholders and their roles and responsibilities in preservation processes. Table 2 provides an idealized categorization of typical stakeholder profiles, based on the SHAMAN reference model, and illustrates each role with the key responsibilities. The exact roles and responsibilities will certainly vary on an organization basis and have to be correspondingly adapted to each specific organizational context. Moreover, additional roles such as technical operators, technology providers and external regulators are not covered here. However, these idealized profiles serve as a guideline for the discussion of responsibilities, expertise and skills. This will provide the background of discussing the role of the *preservation planner*.

Table 2. Roles and their responsibilities.

Title	Role
Executive Manager	This role is responsible for setting the overall goals and objectives of the organization, ensuring that the mandate is fulfilled and the repository continues to serve its designated community. The Executive Manager defines the strategic goals to be achieved and may need to resolve conflicts arising between ends and means.
Responsibilities	
<ul style="list-style-type: none"> • Negotiation and fulfillment of mandate • Assignment of roles and responsibilities in the organization • Strategic planning • Succession planning • Conflict resolution • Organizational and financial management • Financial sustainability • Certification management • Legal compliance 	
Title	Role
Repository Manager	This role is responsible for ensuring repository business continuity, defining business strategies in line with strategic goals, and setting goals and objectives to be achieved by operational management. The Repository Manager operates on the business domain, which requires interaction with the designated communities including producers/depositors and consumers, and the legal environment.
Responsibilities	
<ul style="list-style-type: none"> • Relationship to user communities (producers and consumers) • Relationship to the organizational and legal environment • Awareness of the preservation context • Operational goals and fulfillment • Compliance of business operations with strategic goals 	

<ul style="list-style-type: none"> • Specification of organizational preferences, goals and risks • Authenticity and acceptable loss 	
Title	Role
Technology Manager	The person responsible for technological system continuity and the deployment of technological means to achieve the ends set by the repository business. This role effectively acts as a regulator to the operational manager due to the fact that the choice of technology limits the operational application of means to achieve ends.
Responsibilities	
<ul style="list-style-type: none"> • Technical infrastructure management • IT infrastructure change management • Fulfillment of service level agreements • Operations and reporting • Acquisition of adequate platforms and components 	
Title	Role
Operational Manager	The person responsible for continuous policy-compliant operation of the repository, which involves balancing ends and means and resolving conflicts between them, i.e., constraints as set from Technology Management and Preservation Management. Besides balancing means and ends through decision making, the operational manager is also responsible for overseeing operations and exerting control over operational staff.
Responsibilities	
<ul style="list-style-type: none"> • Content profiling and analysis • Authenticity of content • Operational decision-making to balance ends and means • Making drivers and goals operational • Cost-benefit analysis • Definition of preservation plans • Monitoring of relevant influence factors • Monitoring of the efficient deployment of resources • Compliance monitoring of operations • Controlled experimentation 	
Title	Role
Operator	The person responsible for the operation of the repository and is aware of the details of the design and deployment of the system.
Responsibilities	
<ul style="list-style-type: none"> • Repository operations and monitoring • Preservation operations and monitoring • Execution of experiments • Systems configuration and IT operations • Database management and support • Technical documentation • Reporting 	

From these typical responsibilities, we can see that the role that is responsible for preservation planning is what above is called the *Operational Manager*. However, it is evident that the different roles do not operate independently from each other, but interact on many different levels offering several areas of

friction. Figure 2 (derived from Becker and Rauber, 2011²⁹) shows typical stakeholders and their key relationships within the dimensions *strategic/tactic* and *business/technology* on the left side. It illustrates the problem space and its continuum from technological means to business ends. On the right side, it illustrates the exemplary assignment of these same roles as it is found in the Digital Archive of Austria. It is interesting to note that in this case, three of the management roles are in fact assigned to boards, not single persons. We omit additional related roles and activities such as technology providers and further technical staff.

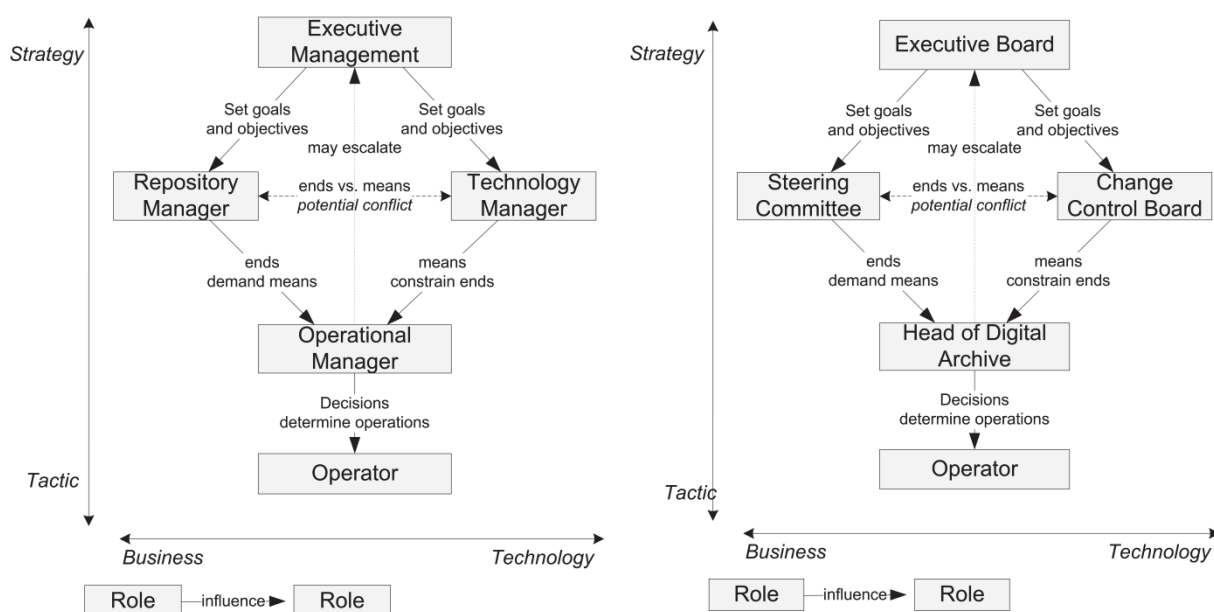


Figure 2. Decision making roles: Idealized stakeholders (left) and exemplary assignment (right).

6. Preservation tasks and required skills and knowledge

Clearly, the actors in a preservation context require a broad set of skills and expertise if they are to preserve digital objects into perpetuity. An understanding of the business goals the organization strives to achieve, the environment it is operating in and the processes it has implemented are just as important as an in-depth knowledge of the technical challenges digital objects pose.

To clarify which specific skills and expertise should be acquired, we build upon the considerable amount of work dedicated to identifying required skills and knowledge for digital curators. The DigCCurr³⁰ project aimed at building a digital curation curriculum. Bähr, Lindlar, and Vlaeminck, 2011³¹ describe necessary know-how and prevalent gaps with a focus on libraries. Engelhardt, Strathmann, &

²⁹ Christoph Becker and Andreas Rauber, "Preservation Decisions: Terms and Conditions apply. Challenges, Misperceptions and Lessons Learned in Preservation Planning," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'11), Ottawa, Ontario, Canada, June 13-17, 2011* (New York, NY: ACM, 2011), 67-76.

³⁰ <http://ils.unc.edu/digccurr> (accessed 31 August 2012)

³¹ Bähr et al., "Puzzling over digital preservation."

McCadden³² report on an international survey on training needs in digital curation and preservation conducted in the DigCurV³³ project.

In the DigCCurr project, a matrix of knowledge and skills necessary for carrying out digital curation work has been developed.³⁴ The authors describe the scope of digital curation and elaborate categories of required knowledge and skills. The matrix identifies and organizes the material that shall be covered in a digital curation curriculum. It is organized along six dimensions: *mandates, value, and principles* (institutional/context specific reasons why the curation functions are carried out and how to evaluate them); *functions and skills* (explicit knowledge of curation methods); *professional, disciplinary, institutional, organizational, or cultural context* (institutional/context specific peculiarities); *type of resource* (as the target of the curation work); *instrumental knowledge* (prerequisite knowledge such as characteristics of technologies); and *transition points in the information continuum* (understanding points of digital content transition from pre-creation to secondary use environments). Each dimension contains categories of knowledge and skills that the authors consider necessary for digital curation work and shall thus be taught to students.

Building upon this, we can position core tasks of preservation planning and analyse the required skills and knowledge. In the following, we list activities necessary for preservation planning and preservation operation and assign to them the categories, as specific areas of skills and knowledge. We add skills and knowledge categories (in **bold**) which we consider important but do not see covered by a category mentioned in Lee and Tibbo, 2011.³⁵ Table 3 gives a description of these additional skills and knowledge categories, The focus hereby is on domain-specific skills that are required by the problem area itself, not on (certainly relevant) management skills such as conflict resolution and coordination. While some of these categories are related to broader categories in Lee and Tibbo, 2011,³⁶ they refer to particular aspects of more specific relevance to preservation than those discussed there. Similarly to the DigCCurr skills matrix, these extended categories are not meant to be exhaustive, but should be understood as identified key areas of expertise that are seen as essential.

Table 4 and Table 5 provide the categories of tasks and skills for preservation planning and preservation operations.

In addition to the core tasks of preservation planning and operations, a number of context-related activities are located in the governance capabilities. These include tasks for managing the preservation context, as shown in Table 6.

As we see from the activities, skills, and expertise areas necessary to successfully plan for digital information longevity, the preservation planner as the person responsible for taking preservation measures needs to have an understanding of both the business goals that need to be achieved and the technical risks that need to be mitigated along that way.

Many decisions taken within the scope of preservation planning are based on contextual information and need to be communicated through appropriate policies. These policies are often related to

³² Claudia Engelhardt, Stefan Strathmann, and Katie McCadden, "Report and analysis of the survey of Training Needs," *DigCurV - Digital Curator Vocational Education Europe Project*, 2011, <http://www.digcur-education.org/eng/Resources/Report-and-analysis-on-the-training-needs-survey>.

³³ <http://www.digcur-education.org/>

³⁴ Christopher A. Lee and Helen Tibbo, "Where's the Archivist in Digital Curation? Exploring the Possibilities through a Matrix of Knowledge and Skills," *Archivaria* 72 (Fall 2011): 123-168.

³⁵ Ibid.

³⁶ Ibid.

Table 3. Knowledge and skills.

Knowledge and skills category	Explanation
Content Analysis	Understanding of the characteristics of digital content, digital encodings and formats. This includes format risks, content corruption, interoperability between environments, dependencies, format identification methods and tools as well as scalable analysis of large content collections.
Requirements Analysis	Knowledge and identification of the criteria categories relevant for the planning scenario. Identification and bringing together of relevant stakeholders around a table to gather requirements. Guidance of requirements elicitation workshops with a focus on steering participants towards an accurate problem description rather than anticipating solutions.
Preservation Actions	Understanding and appreciation of the range of potential preservation actions including a differentiation according to type. No preservation action type (such as migration) should be given priority to another (such as emulation) without knowing their strengths and weaknesses in a certain application domain.
Preservation Action Quality Assurance	Development and implementation of quality assurance workflows especially with respect to automation. Understanding possible weaknesses of software programs. Measurement and correct interpretation of criteria necessary to assure a digital object's quality.
Information Sources, Evidence, and Trustworthiness	Understanding of the role of evidence for trustworthiness. Knowledge of the kinds of sources of information that can provide indicators and evidence for decisions. Knowing knowledge bases from where measurements to certain properties can be obtained from and assessing their trustworthiness.
Controlled Experimentation and measurement	Knowledge of how properties of digital objects, formats, software, software executions can be measured. Appreciating the importance of the usage of relative measures instead of absolute to facilitate comparison. Understanding of experiment design, execution and documentation of evidence. Recognizing the importance of replicability and repeatability of experiment results in the long term as an essential part of evidence-based decision making.
IT Operations, Quality, and Management	Understanding of the basic acquisition, development, evaluation, configuration, operation, maintenance, and change management processes required to successfully operate IT systems. This includes an understanding of the concept of software quality as defined for example in ISO 25010 'Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models' (International Standards Organization 2011)
Decision making with multiple criteria	Understanding of approaches to quantify and compare multiple criteria and take accountable decisions in a systematic way. This includes trade-offs between conflicting objectives, and the ability to make goals and constraints (such as costs and benefits) operational
Risk Management	Understanding of the nature of uncertainty and risk management and its relevance in the preservation context. This includes concepts such as uncertainty, probability, likelihood, impact, mitigation, opportunities, risk assessment methods, and continuous risk monitoring.

the regulatory environment or to compliance criteria such as those contained in the ISO 16363 Repository Audit and Certification catalogue (International Standards Organization 2012). Preservation planners have to be able to assess their organizational implications and understand how and where they enter the

Table 4. Required skills and knowledge for preservation planning tasks.

Core tasks of preservation planning	Required skills and knowledge
Make drivers and goals operational, i.e., define objectives and constraints represented by decision criteria.	Significant Properties; Controlled Experimentation and measurement ; Analysis and Characterization of Digital Objects/Packages; Automation; IT Operations, Quality, and Management ; Requirements Analysis ; Preservation Actions ; Preservation Action Quality Assurance ; Information Sources, Evidence, and Trustworthiness ; Decision making with multiple criteria
Select a (minimal) set of relevant preservation actions for assessment which potentially fulfil the defined requirements.	Automation; Archival Storage; Common Services; Scale and Scalability; Purchasing and Managing Licenses to Resources; Transformation of Digital Objects/Packages; Preservation Actions
Assess preservation actions against the specified requirements.	Automation; Analysis and Characterization of Digital Objects/Packages; Information Sources, Evidence, and Trustworthiness ; Controlled Experimentation and measurement
Specify actions and directives in an understandable form and deliver it to the unit responsible for deployment.	Automation; Archival Storage; Common Services; Preservation Action Quality Assurance

Table 5. Required skills and knowledge for preservation operation tasks.

Core Tasks of Preservation Operation	Required skills and knowledge
Execution of preservation actions according to preservation plans and ensure full documentation of the execution process.	Evidence; Provenance; Scale and Scalability; Robustness; Description, Organization, and Intellectual Control; Transformation of Digital Objects/Packages; Transfer; IT Operations, Quality, and Management
Analysis of technical characteristics of content	Significant Properties; Provenance; Stakeholders; Analysis and Characterization of Digital Objects/Packages; Characteristics of Information and Record Creating Environments; Level of Aggregation; Characteristics of Technologies; Content Analysis ; Risk Management ; IT Operations, Quality, and Management ; Preservation Actions ; Controlled Experimentation and measurement
Control the processing of appropriate preservation metadata corresponding to chosen standards.	Evidence; Provenance; Description, Organization, and Intellectual Control; IT Operations, Quality, and Management
Measure properties of renderings/performances and compare them to each other to measure their equivalence corresponding to requirements.	Significant Properties; Description, Organization, and Intellectual Control; Validation and Quality Control of Digital Objects/Packages; Preservation Action Quality Assurance ; IT Operations, Quality, and Management
Produce documentation of activities in an adequate and understandable form.	Evidence; Provenance; Description, Organization, and Intellectual Control; IT Operations, Quality, and Management

Table 6. Required skills and knowledge for managing preservation context.

Core Tasks of Managing the Preservation Context	Required skills and knowledge
Collect and describe all relevant influence factors that facilitate or restrict the decision for a preservation action; i.e., all drivers, constraints and goals applicable. Examples for external influencers include user communities, access requirements or regulations. ³⁷	Accountability; Authenticity; Chain of custody; Context; Provenance; Significant properties; Stakeholders; Standardization; Sustainability; Trust; Transformation of Digital Objects/Packages; Professional Contexts; Disciplinary Contexts; Institutional or Organizational Contexts; Characteristics of Information and Record Creating Environments; Format; Requirements Analysis; Information Sources, Evidence, and Trustworthiness
Define the potential communities of users to the organization's holdings and document their requirements and available access means.	Context; Stakeholders, Sustainability; Trust; Collaboration, Coordination, and Contracting with External Actors; Professional Contexts; Disciplinary Contexts; Institutional or Organizational Contexts; Characteristics of Information and Record Creating Environments
Monitor internal and external influence factors of relevance.	Long Term, Open Architecture, Stakeholders, Analysis and Characterization of Digital Objects/Packages; Professional Contexts; Disciplinary Contexts; Preservation Actions; Risk Management; IT Operations, Quality, and Management
Analyse and document preservation and access requirements in the organization's communities.	Stakeholders; Significant Properties; Context; Sustainability; Trust; Collaboration, Coordination, and Contracting with External Actors; Institutional or Organizational Contexts; Characteristics of Information and Record Creating Environments; Requirements Analysis

planning process. It is the preservation planner's duty to be aware of and understand the organization's policies and ultimately adhere to them in planning so that the specified operations are policy-compliant. However, current methods and tools provide only incomplete specifications and means to support this coordination between context-awareness, guidance and strategies, and operational decision making.

The preservation planner must have detailed knowledge of digital objects relevant to the organization, be it ordinary files or database records, and their rendering through the characteristics that must be preserved to maintain its authenticity and understandability. Preservation actions which appear as software tools depend on computer hardware, operating systems and software libraries and need to be understood with regard to their long-term suitability. Finally, the preservation planner also needs to provide crucial input to the mandate negotiation process, since it is the planner's responsibility to answer the question whether a particular digital object's authenticity and understandability can be preserved with the available means.

A recent study of training needs in digital preservation revealed that respondents indicate 'Preservation & data management planning' and 'Preservation tools' amongst the areas where they are lacking the required knowledge and skills the most.³⁸ This may also come from the yet little-understood planner role including its knowledge and skills required to be capable of preserving the complex

³⁷ A comprehensive list of internal and external drivers is given in SHAMAN 2012.

³⁸ Engelhardt et al., "Report and analysis of the survey of Training Needs."

dependency network of file formats, software, operating systems, and hardware. Irrespective of the actual profession of the person planning for preservation – be it an archivist, librarian or digital curator – we showed that solid IT knowledge is prerequisite to successful preservation planning. This needs to be combined with skills in organizational understanding, as for instance imparted in business informatics studies. Well-established standards and methods from Information Systems and Software Technology need to be followed in the operational planning of preservation measures.

Conclusions and Outlook

In this article, we discussed the current state of the art in decision making for digital preservation planning and operations and elaborated on required activities. Effective governance requires a clear assignment of roles and responsibilities. Based on a socio-technical perspective on the capabilities required for ensuring digital information longevity, we illustrated typical roles and their responsibilities in the areas of Preservation Planning and Preservation Operations. We listed tasks required to prepare a preservation plan, and associated each task to the knowledge and skills that the responsible person should possess. We furthermore highlighted that an ideal preservation planner combines solid IT knowledge and skills with an understanding of organizational processes.

An organization that intends to develop systematic preservation planning and operations capabilities requires a well-defined governance framework and methods for diagnosing existing capabilities and defining a roadmap for capability development. Although widely accepted, catalogues of compliance criteria such as ISO 16363 do not support organizations in systematically assessing and improving their capabilities. In general, the wide body of digital preservation reference models and frameworks provide a common language, building blocks, and other types of knowledge derived from an in-depth analysis of the domain. However, these models are not always well-founded and consistent. Systematic approaches for governance can be adopted from fields of Information Systems and IT Management.³⁹

While this article focused on preservation planning and operations, it is clear that managing the preservation context requires a similarly systematic approach. On the one hand, the corresponding capabilities and governance processes need to be clearly specified. On the other hand, “preservation policies” are defined on very different levels of granularity, clarity, and ambiguity. There is a strong need for a well-defined, standard approach to representing organizational preservation goals, objectives, constraints and directives in a systematic way to ensure that the preservation context is properly documented and communicated. This is also a crucial enabler for increased automation in preservation planning and operations.

Acknowledgements

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

³⁹ Christoph Becker, Gonçalo Antunes, José Barateiro, Ricardo Vieira, and José Borbinha, “Control Objectives for DP: Digital Preservation as an Integrated Part of IT Governance,” in *Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology (ASIST 2011)*, New Orleans, LA, USA, October 9-13, 2011, http://publik.tuwien.ac.at/files/PubDat_203334.pdf.

Bibliography

- Bähr, Thomas, Michelle Lindlar, and Sven Vlaeminck. "Puzzling over digital preservation – Identifying traditional and new skills needed for digital preservation." *World Library and Information Congress: 77th IFLA General Conference and Assembly*. 2011.
- Becker, Christoph, and Andreas Rauber. "Preservation Decisions: Terms and Conditions apply. Challenges, Misperceptions and Lessons Learned in Preservation Planning." In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'11), Ottawa, Ontario, Canada, June 13-17, 2011*, 67-76. New York, NY: ACM, 2011.
- Becker, Christoph, Gonçalo Antunes, José Barateiro, and Ricardo Vieira. "A Capability Model for Digital Preservation: Analyzing Concerns, Drivers, Constraints, Capabilities and Maturities." In *Proceedings of the 8th International Conference on the Preservation of Digital Objects (iPRES2011), Singapore, November 1-4, 2011*, 1-10. Singapore: National Library Board and Nanyang Technological University, 2011.
- Becker, Christoph, Gonçalo Antunes, José Barateiro, Ricardo Vieira, and José Borbinha. "Control Objectives for DP: Digital Preservation as an Integrated Part of IT Governance." In *Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology (ASIST 2011), New Orleans, LA, USA, October 9-13, 2011*. http://publik.tuwien.ac.at/files/PubDat_203334.pdf.
- Becker, Christoph, Hannes Kulovits, Mark Gутtenbrunner, Stephan Strodl, Andreas Rauber and Hans Hofman. "Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans." *International Journal on Digital Libraries* 10, no. 4 (December 2009): 133-157.
- Brown, Richard Harvey, and Beth Davis-Brown. "The making of memory: the politics of archives, libraries and museums in the construction of national consciousness." *History of Human Sciences* 11, no. 4 (1998): 17-32.
- Carr, Edward Hallett. *What Is History?* New York: Random House, 1961.
- Consultative Committee for Space Data Systems. "Reference Model for an Open Archival Information System (OAIS)." 650.0-M-2, 2012.
- Cox, Richard J. "The Documentation Strategy and Archival Appraisal Principles: A Different Perspective." *Archivaria* 38 (1994): 11-36.
- CRL, and OCLC. "Trustworthy Repositories Audit Certification: Criteria and Checklist (TRAC)." The Center for Research Libraries (CRL) and Online Computer Library Center (OCLC), 2007.
- Diamond, Elizabeth. "The Archivist as Forensic Scientist -- Seeing Ourselves in a Different Way." *Archivaria* 38 (1994): 139-154.
- Duranti, Luciana, and Barbara Endicott-Popovsky. "Digital Records Forensics: A New Science and Academic Program for Forensic Readiness." *Journal of Digital Forensics, Security and Law* 5, no. 2 (2010): 45-62.
- Engelhardt, Claudia, Stefan Strathmann, and Katie McCadden. *Report and analysis of the survey of Training Needs*. Digital Curator Vocational Education Europe Project, 2011.
- Formats - RealNetworks, Inc. *Supported Media Formats*. http://cache-download.real.com/free/windows/mrkt/help/RealPlayer-15/en/Content/Media_Types.htm.
- Gardner, John. "The Mark of Responsibility." *Oxford Journal of Legal Studies* 23, no. 2 (2003): 157-171.

- Halbwachs, Maurice. *On Collective Memory* (originally published in 1941, edited and translated by Lewis A. Coser). Chicago: University of Chicago Press, 1992.
- Hedstrom, Margaret. "Exploring the Concept of Temporal Interoperability as a Framework for Digital Preservation." *Third DELOS Network of Excellence Workshop on Interoperability and Mediation in Heterogeneous Digital Libraries*. 2001.
- International Standards Organization. "ISO 14721:2012." *Reference model for an open archival information system (OAIS)*. Consultative Committee on Space Data Systems, 2012.
- International Standards Organization. *Space data and information transfer systems -- Audit and certification of trustworthy digital repositories (ISO 16363:2012)*. ISO, 2012.
- . "Systems and software engineering – System and software Quality Requirements and Evaluation (SQuARE) – System and software quality models (ISO/IEC 25010)." 2011.
- IT Governance Institute. "COBIT 5 – A business Framework for the Governance and Management of Enterprise IT. 2012." 2012.
- Josias, Anthea. "Toward an understanding of archives as a feature of collective memory." *Archival Science* 11, no. 1-2 (2011): 95-112.
- Kulovits, Hannes, Andreas Rauber, Markus Brantl, Astrid Schoger, Tobias Beinert, and Anna Kugler. "From TIFF to JPEG2000? Preservation planning at the Bavarian State Library using a collection of digitized 16th century printings." *D-Lib Magazine* 15, no. 11/12 (November/December 2009). <http://www.dlib.org/dlib/november09/kulovits/11kulovits.html>.
- Lee, Christopher A., and Helen R. Tibbo. "Where's the Archivist in Digital Curation? Exploring the Possibilities through a Matrix of Knowledge and Skills." *Archivaria* 72 (Fall 2011): 123-168.
- Liu, Wayne W. "Identifying and Addressing Rogue Servers in Countering Internet Email Misuse." In *IEEE/SADFE-2010: Proceedings of the 5th International Workshop on Systematic Approaches to Digital Forensic Engineering*, 13-24. Oakland, CA: IEEE Computer Society, 2010.
- . "Trust Management and Accountability for Internet Security." Ph.D. diss. Florida State University, July 2011.
- Liu, Wayne, Sudhir Aggarwal, and Zhenhai Duan. "Incorporating Accountability into Internet Email." In *SAC'09: Proceedings of the 24th ACM Symposium on Applied Computing*. Honolulu, HI: ACM, 2009.
- Long, Andrew Stawowczyk. "Long-Term Preservation of Web Archives – Experimenting with Emulation and Migration Methodologies." International Internet Preservation Consortium (IIPC), 10 December 2009. Accessed 1 April 2012. <http://www.netpreserve.org/sites/default/files/resources/Methodologies.pdf>.
- Lord, Philip, and Alison Macdonald. "e-Science Curation Report: Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision." The JISC Committee for the Support of Research (JCSR), 2003.
- Misztal, Barbara. *Theories of Social Remembering*. Maidenhead, UK: Open University Press, 2003.
- MPlayer. <http://www.mplayerhq.hu/>.
- Nora, Pierre. "Between Memory and History: Les Lieux de Memoire." *Representations* 26, Special Issue (Spring 1989): 7-24.
- Object Management Group. "Business Motivation Model 1.1." 2010.

- Pruzan, Peter. "From Control to Values-Based Management and Accountability." *Journal of Business Ethics* 17, no. 13 (October 1998): 1379-1394.
- Rau, Kenneth G. "Effective Governance of IT: Design objectives, roles, and relationships." *Information Systems Management* 21, no. 4 (2004): 35-42.
- Real Networks. *RealNetworks, Inc. - Acquisition History*. Accessed 5 July 2012.
<http://investor.realnetworks.com/faq.cfm?faqid=2>.
- RealNetworks, Inc. "Download the VivoActive Player." <http://egg.real.com/vivo-player/vivodl.html/>.
- Rosenthal, David S. H., Thomas Lipkis, Thomas S. Robertson, and Seth Morabito. "Transparent Format Migration of Preserved Web Content." *www.dlib.org*. Jan 1, 2005. Accessed 5 July 2012.
<http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html>.
- Sachs, Albie. "Archives, truth and reconciliation." *Archivaria* 62 (2006): 1-14.
- Schwartz, Joan M., and Terry Cook. "Archives, records, and power: the making of modern memory." *Archival Science* 2, no. 1 (2002): 1-19.
- SHAMAN. *SHAMAN Reference Architecture v3.0*. Deliverable, SHAMAN project, 2012.
- Sinclair, Amanda. "The Chameleon of Accountability: Forms and Discourses." *Accounting, Organizations and Society* 20, no. 2/3 (1995): 219-237.
- Smith, MacKenzie, and Reagan W. Moore. "Digital Archive Policies and Trusted Digital Repositories." *International Journal of Digital Curation* 1, no. 2 (2007).
- Wallace, David A. "Introduction: memory ethics--or the presence of the past in the present." *Archival Science* 11 (2011): 1-12.
- Wallace, David A., and Lance Stuchell. "Understanding the 9/11 Commission Archive: Control, Access, and the Politics of Manipulation." *Archival Science* 11, no. 1-2 (2011): 125-169.
- Webb, Colin. "Guidelines for the Preservation of Digital Heritage." Prepared by the National Library of Australia for the Information Society Division, UNESCO, March 2003.
<http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>.
- Yakel, E., P. Conway, M. Hedstrom, and D. Wallace. "Digital Curation for Digital Natives." *Journal of Education for Library and Information Science* 55, no. 1 (2011): 23-31.
- Yakel, Elizabeth. "Digital curation." *OCLC Systems & Services: International digital library perspectives* 23, no. 4 (2007): 335-340.

Posters and Presentations

Challenges and Opportunities of Digitizing and Preserving Cultural Heritage in Zimbabwe

Collence Takaingenhamo Chisita¹ and Amos Bishi²

¹ School of Information Sciences, Harare Polytechnic, collencechisita@yahoo.com

² School of Information Sciences, Harare Polytechnic, bishyy85@cooltoad.com

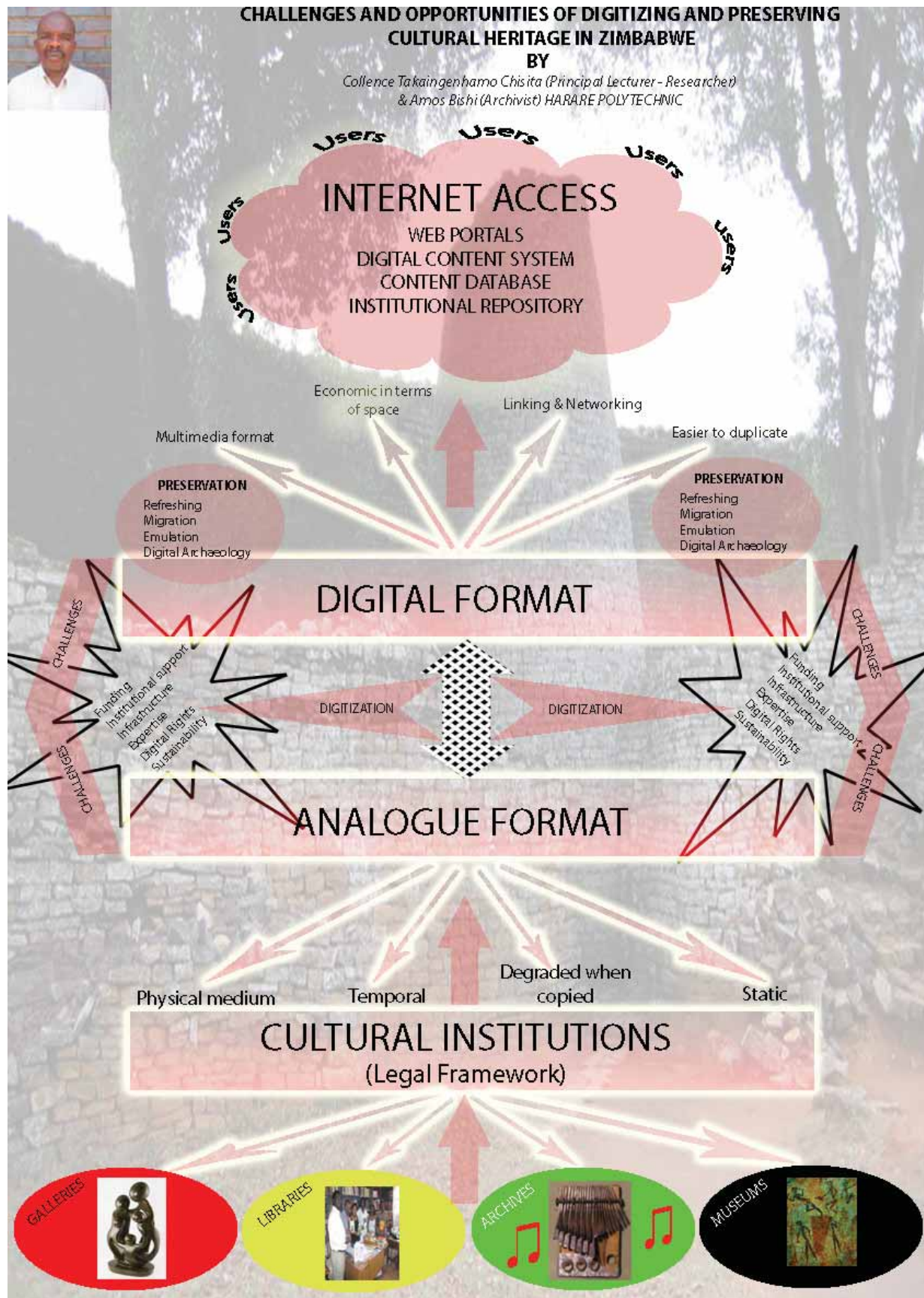
Abstract

The Information and Communication Technology (ICT) revolution has transformed the creation, use, and preservation of cultural heritage in Zimbabwe. The pervasive nature of technology makes it imperative for the Cultural heritage institutions in Zimbabwe to place digitization and preservation at the heart of their strategic plans. Zimbabwe has a rich cultural heritage, which should be digitized and preserved as the country joins the global knowledge economy since we are now operating in the post-modernism technology-driven environment. In this poster, the authors explore what needs to be done in order to ensure a successful and sustainable digitization and preservation programme in cultural heritage institutions. The poster highlights the strategies that are in place to ensure the digitization and preservation of documents as well as the challenges and opportunities of migrating cultural heritage from analogue to digital format and the factors that impede the digitization and preservation of cultural heritage.

Authors

Collence T. Chisita is a Principal Lecturer in the School of Information Sciences at Harare Polytechnic. He holds an Msc. in Information Science and is currently completing a Ph.D. in Information Sciences. Currently he lectures in Records/Information Consultancy and Brokerage, Sociology of Information, among other subjects. He is a member of Records and Archives Information Management of Zimbabwe and other international organisations like IFLA and Euro Africa forum for ICT Research. He has worked extensively with the National Archives of Zimbabwe.

Amos Bishi is a Lecturer in Records and Archival Management and is based at Harare Polytechnic's School of Information Sciences. He is also a part-time lecturer at the Zimbabwe Open University's School of Information Sciences. He holds a Bachelor of Science Honors Degree in Information Science. He is currently studying for a Master's degree in Records Management. He has also provided services to the Africa Capacity Building Foundation as a senior project Digital Asset Archivist.



The long and winding road from aspiration to implementation: building an enterprise digitisation capability at the University of Melbourne

Donna McRostie

University Library

Abstract

The University of Melbourne has embarked upon a bold strategy to ensure its place as one of the finest universities in the world, a strategy founded upon excellence in research, learning and teaching, and engagement. Supporting this vision and released in July 2008, Melbourne's Scholarly Information Future is a ten-year strategy to guide the development of the University's scholarly information services, collections, systems, technologies and infrastructure. This strategy sits alongside research, e-learning, engagement and other domain-specific plans supporting the Growing Esteem agenda.

The University of Melbourne has a rich, complex, and ultimately voluminous, array of cultural, scholarly and research material that is of great interest and value to the University community, scholarly researchers and the global community. The strategy identified in its aspirations the importance of building effective access to the rich cultural, scholarly and research collections of the University and acknowledges the critical role that digitisation plays in achieving this vision. Typically in 2008 text and object based digitisation throughout the University was carried out in an uncoordinated, ad hoc manner. There was also a lack of centralised expertise which had led to a proliferation of isolated, under-resourced areas producing inconsistent and indifferent quality images in the absence of well-documented enterprise wide digitisation standards.

This led to the prioritisation and investment from the University Library to develop an exemplar digitisation framework, program and enterprise capability for the University to leverage.

Four years later this is still a work in progress and this paper will review our journey and the approach we took in building this capability in a challenging economic environment, the engagement strategies to gain support and funding, skills and equipment as well as the unique challenges of the digitisation of a diverse array of University collections.

Authors

Donna is a professionally qualified Archivist and Records Manager and holds a Graduate Diploma in Information Management and Graduate Certificate of University Management. Donna has worked in a variety of Information Management roles in higher education over the last 15 years and in her current role Donna is responsible for leading and managing the Information Management Program, including building an enterprise digitization capability in the University Library and supporting the University Librarian on the delivery of the University's Scholarly Information Strategy.



*The long and winding road from aspiration to implementation;
building an enterprise digitisation capability at
The University of Melbourne.*

Donna McRostie

Director, Information Management

University Library

BUILDING AN ENTERPRISE CAPABILITY

In 2008 the Vice Chancellor led an Information Futures Commission which, after broad consultation, developed **Melbourne's Scholarly Information Future**, a ten-year strategy that guides scholarly information services, collections, systems, technologies and infrastructure.

A key aspiration of this strategy is the imperative to unlock the rich, complex array of cultural materials of great interest and value to scholars as well as the global community. The University acknowledges the critical role that digitisation plays in achieving this vision.

HOW we got established

- Governance: Policy, Strategy, Road Map, Advisory Committee
- Road Map linking closely to strategy resulting in central funding for purpose built facility
- University funds invested to support the Scholarly Information Future aspirations together with Library investments enabled purchase of key equipment to increase scope of capability and output
- Identification of digitisation capability as a key research infrastructure platform for the University



HOW we developed the expertise

- Built on established capability grounded in microform services and corporate scanning
- Leveraged expertise of the Library's eScholarship Research Centre with a world reputation in digital scholarship in contemporary research practice
- Scaled up the range of equipment with modest investment in additional staff
- Expanded service scope – scanning services, consultancy and advice, training & outreach
- Not attempting to undertake all digitisation for the University but acts as the primary means through which standardized and sustainable practices are promulgated



Establishment Challenges

- **Demand outstrips capacity** – established self service facilities for the University community, prioritising of projects + outsourcing options
- **Skills** – mostly developed "on the job" with foundation in photography, micrographic and imaging services
- **Technical challenges** – unique challenges required specialists technical skills not met by central IT services – translating manual process to more integrated and automated model of scanning methods and workflows with flexibility and adaptability to suit unique requirements of each job
- **Resourcing** – skills transfer to other library staff through professional secondment opportunities; externally funded projects build pool of casual staff with University students prime targets
- **Sustainability** – maintain profile to ensure capability is considered as key research infrastructure and leverage funding opportunities; income generation – provision of services to external cultural institutions
- **Digital Preservation** – lag time between the development of the capability and a robust curation and preservation processes.

University Digitisation Service
The University of Melbourne Library, Parkville, Victoria 3010 Australia
P: +61 3 8344 6163 E: digitisation-enquiries@unimelb.edu.au

DIGITISATION AS TRANSFORMATION

Digital surrogates of cultural resources can transform cultural collections

- Accessibility
- Computational manipulation
- Visualisation in new and many places
- Content analysis

Marketing/engagement tool

- Showcasing treasures
- Attracting benefactors
- Reaching new markets and communities

Connecting people and heritage

- Creating, preserving and connecting public knowledge and communities
- Protect the original artefact



Google Earth overlay



University of Melbourne Open Library book reader



Scanned at higher resolutions to preserve fine detail



Lighting for illuminated manuscripts used controlled reflection to show the extent of gold leaf

<http://digitisation.unimelb.edu.au>

Retrieving a part of Danish colonial history

From dust to digital copy

Asger Svane-Knudsen and Jiří Vnouček

Authors

Asger Svane-Knudsen graduated in 1987 from the University of Southern Denmark and the University of Copenhagen with a Master's in history and geography. After six years as a consultant with an IT company, he became archivist at the Danish National Archives in 1995 and, from 2005, curator at the Danish State Archives, with responsibility for preservation and conservation, where he initiated a project about conservation, preservation and copying files from Danish Overseas Trading Companies.

Jiří Vnouček graduated in 2010 from The Royal Danish Academy of Fine Arts – School of Conservation with a Master's of Conservation-restoration Science. After more than 25 years of experience as a conservator of books and rare documents, he is now employed at the Preservation department of the Royal Library in Copenhagen as conservator of paper, parchment and books. From 2008, he has been working on the project of the conservation of protocols from Trankebar, part of the Danish Overseas Trading Company's Archive from the Danish National Archive

Poster presentation at the UNESCO conference "The Memory of the World in the Digital Age: Digitization and Preservation",
26-28 September 2012, Vancouver, Canada

Retrieving a part of Danish colonial history - from dust to digital copy



STATENS ARKIVER

Asger Svane-Knudsen
Danish State Archive, Copenhagen
ask@sa.dk

Jifi Vnouček
The Royal Library, Copenhagen
jiv@kb.dk



Retrieving a part of Danish colonial history – from dust to digital copy

Slide 2 of 9

Archives of the Danish Overseas Trading Companies

The Danish tropical colonies were managed by the overseas trading companies, such as Danish East India Company, the Danish Asiatic Company, the Danish West India and Guinea Company and the Danish West India Trading Company. They are from the period 1620-1917.

The archives of these companies were created in both temperate climate in the head offices in Copenhagen and in tropical climate in the colonies. Many of the documents from the tropical colonies are in a very deteriorated condition.

Danish National Archives holds all existing archives, and they consist of around 4000 protocols and boxes with royal charters, registers, reports, copybooks, letters, instructions, accounts, ship's logs, maps etc.

This cultural heritage is important for the history of Denmark, the history of world colonialism and the history of the nations around the colonies.



The merchant ship "Dronning Juliane Marie" around 1765.
Photo: Danish Maritime Museum



The connection between the colonies and Denmark was only by ship.
The archives created in the colonies were sailed to Copenhagen, at the time when the colonies were ceded.



The archives of the Danish Overseas Trading Companies were included in the Memory of the World Register in 1997



STATENS ARKIVER

Asger Svane-Knudsen
Danish State Archive, Copenhagen
ask@sa.dk

Jifi Vnouček
The Royal Library, Copenhagen
jiv@kb.dk



Retrieving a part of Danish colonial history – from dust to digital copy

Slide 3 of 9

Danish Tropical Colonies

Christiansted, St. Croix

Fortress Prinsensten, Guinea

Fortress Dansborg, Trankebar

Danish Church, Serampore

STATENS ARKIVER

Asger Svane-Knudsen
Danish State Archive, Copenhagen
ask@sa.dk

Jiří Vnouček
The Royal Library, Copenhagen
jiv@kb.dk

Retrieving a part of Danish colonial history – from dust to digital copy

Slide 4 of 9

Problems with the access to damaged documents

Unfortunately the physical condition of some documents from the archives is in such a bad state, that researchers have been denied access to them. This is a rather complicated situation as many of the documents are not possible to derive any information from, they are simply too deteriorated and almost slowly turning into dust.



Condition of the documents

For an example the protocols from Trankebar, which are written with iron-gall ink on handmade paper, are heavily damaged due to the tropical climate in East India. The sad state of these documents is a result of a combination of several deterioration processes: iron gall ink corrosion, paper fiber degradation, and insect and mould damage. Many records are written on paper which is very fragile and have darkened to such an extent that the text is almost impossible to read.



STATENS ARKIVER

Asger Svane-Knudsen
Danish State Archive, Copenhagen
ask@sa.dk

Jiří Vnouček
The Royal Library, Copenhagen
jiv@kb.dk

Retrieving a part of Danish colonial history – from dust to digital copy

Slide 5 of 9

History of preservation and conservation



Silk textile repair



New cardboard binding

In the last half of the 20th century several attempts to rescue the physical documents were undertaken, but the conservation treatments brought rather controversial results. Meanwhile some parts of the text of the records were saved, others were lost forever. The conservation project was interrupted leaving hundreds of untreated documents in boxes, waiting for their slow decay.

Old methods of stabilisation



Neutralization by Barrow method



Japanese paper repair



Repair by hot lamination tissue



STATENS ARKIVER

Asger Svane-Knudsen
Danish State Archive, Copenhagen
ask@sa.dk

Jiří Vnouček
The Royal Library, Copenhagen
jiv@kb.dk



Retrieving a part of Danish colonial history – from dust to digital copy

Slide 6 of 9

New methods and current conservation treatment



In situ repairs of cracks by Japanese paper

In 2008, the Preservation department of the Royal Library in Copenhagen resumed conservation of these archival records kept at the Danish National Archive. The results and recommendations made by an international group of researchers studying the processes of iron gall ink degradation were found to be very promising <http://ink-corrosion.org/>. Based on these results it was decided on a complex conservation treatment of the documents, including chemical stabilization of the ink as well as different kinds of paper repairs. The main target was to restore the physical condition of the documents so they could be preserved, safely handled and later digitized. It was possible to improve most of the documents but a group of material with very heavy damages due to ink corrosion and paper deterioration, had to be treated in another way.



Comparison of number of cracks before and after phytate treatment



Protocol from Trankebar from 1697-1698 before and after leafcasting



STATENS ARKIVER

Asger Svane-Knudsen
Danish State Archive, Copenhagen
ask@sa.dk

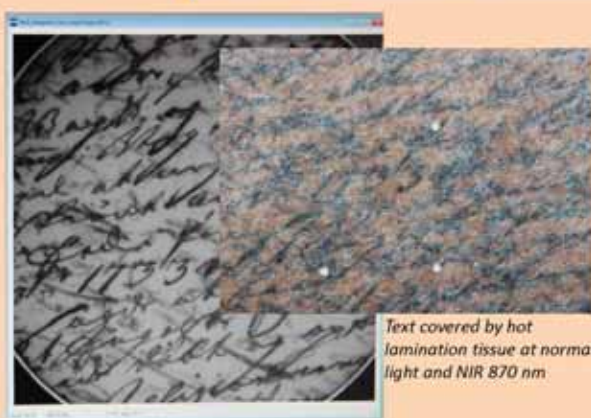
Jiří Vnouček
The Royal Library, Copenhagen
jiv@kb.dk



Retrieving a part of Danish colonial history – from dust to digital copy

Slide 7 of 9

Change of preservation and conservation strategies



Text covered by hot lamination tissue at normal light and NIR 870 nm



Comparison of readability of text in normal and UV light (blue)

Special photography – new chance for hardly visible text

The conservation strategy was re-evaluated and it was decided that the conservation process had to be changed for a group of heavily damaged documents. Instead of improving the physical condition of the documents, prior to the digitization, these documents will be digitized untreated but with the direct assistance of a conservator providing immediate first aid conservation repairs allowing the pages to be turned and securing that all the loose parts of the documents are made visible. To improve the readability of the hardly visible text, multispectral imaging will be used during digitization.



Asger Svane-Knudsen
Danish State Archive, Copenhagen
ask@sa.dk

Jiří Vnouček
The Royal Library, Copenhagen
jiv@kb.dk



Retrieving a part of Danish colonial history – from dust to digital copy

Slide 8 of 9

Digital preservation



Testing VideometerLab, a danish tool for digital copying, using 18 different predetermined wavelengths of light in UV and NIR range.

Digitization of each document

- The digital copy is the only future access
- Must show as many aspects as possible
- Readability improved with UV and infrared light
- Pages digitized with dozens of pictures representing each installation
- Pages might only be digitized sectional due to the equipment
- No manipulated pictures

Two 17th century documents in front of a rack with many terabytes of harddisk space.



The digital copies

- Preserved in an electronic records management system
- Pictures in format for digital archiving (tiff or jpeg2000)
- Registration of all necessary information in the metadata
Metadata consist of several levels, e.g.:
 - General (equipment setup, date, operator, archival unit etc)
 - Document (type, reg.no., description, years, order, extent etc)
 - Page (page no, condition, sectional division etc)
 - Picture (section id, wavelength, angle and distance of light etc)



Asger Svane-Knudsen
Danish State Archive, Copenhagen
ask@sa.dk

Jiří Vnouček
The Royal Library, Copenhagen
jiv@kb.dk



Retrieving a part of Danish colonial history – from dust to digital copy

Slide 9 of 9

Acknowledgement

Contact to the authors:

Asger Svane-Knudsen, ask@sa.dk

Archivist, MA in history and geography, The Danish State Archive, Copenhagen

Jiří Vnouček, jiv@kb.dk

Conservator MSc, The Royal Library, Copenhagen

Conservators:

The Royal Library, Copenhagen:

Henriette Berg, Hanne Karin Sørensen, Jiří Vnouček, Annemarie Teuber

Special thanks:

Matthew James Driscoll and Michael Lerche Nielsen from Arnamagnean Institute, Copenhagen

Jens Michael Carstensen and Karsten Hartelius, Videometer Inc., Lyngby, Denmark

Erik Gøbel, Danish State Archive, Copenhagen

Photos:

If not stated otherwise, courtesy of the Royal Library and the Danish State Archives



STATENS ARKIVER

Asger Svane-Knudsen
Danish State Archive, Copenhagen
ask@sa.dk

Jiří Vnouček
The Royal Library, Copenhagen
jiv@kb.dk



A Paradigm for the preservation of national digital memory of Iran

Mitra Samiee and Saeed Rezaei Sharifabadi

Abstract

Digital preservation consists of two steps: safe storage and permanent access. To ensure safe-keeping of digital objects, a digital archive dedicated to long-term maintenance is needed. Such an archive preserves the bit-stream and its format and metadata is registered to ensure description and retrieval. A proper familiarity with the characteristics of the resources and also possible threats of permanent access to the resources in question is what really matters and should therefore be taken into account in order to attain this goal. The digital storage media are qualitatively vulnerable and thus a precise procedure should be developed to back up and refresh software, hardware and new information carriers. The future rendering of the stored digital items is another issue. How can we be sure we can view and execute digital objects in ten, fifty or even a hundred years from now? Strategies such as migration, conversion, normalization and emulation should be considered and developed. But we also need tools that ensure this permanent access.

Thus, there are fewer opportunities for us to take decisions on the preservation of resources compared with written resources and so we should act immediately to preserve the resources before they fall into disuse. In view of the fact that storage technologies of digital information are rapidly developing and that everyday new types of these technologies are being brought to the market, the old media consequently fall into disuse and subsequent to that, the information dependent on these obsolete technologies quickly become inaccessible. So, the biggest challenge in digital preservation is technological obsolescence that affects not only the migration of the data itself, but also the migration and emulation of technological platform. Despite striking technical achievements as regards the speed of processing and storage capacity, the preservation of digital material will go further than accumulation of information on a high capacity disc and again its retrieval. The problem is that we should be able to make the stored information accessible rather than making this digital information just presentable. Furthermore, we should be able to keep them alive and accessible for the next generations and at the same time maintain the coherence and homogeneity of information as well as the intellectual property rights and copyrights of their owners. From 2000 onwards, many of digital national libraries in the world, including the American Library of Congress, have developed an initiative for the establishment of an appropriate infrastructure with a view to managing and preserving their valuable works in digital format. The American Library of Congress, within the framework of NDIIPP (The National Information Infrastructure and Preservation Program), the National Library of Australia within the framework of ICABS's third clause concerning the collection and preservation of digital resources and Netherland's National Library by means of DIAS (Digital Information Archival System) and other National libraries through other various programmes and projects have dealt with the preservation of their valuable digitized resources.

Similarly in Iran, since 2001, some initiatives have been adopted by the National Library and Archive of Republic of Iran to digitize the library resources and the development of a digital library has been officially proposed. To begin with, a working group consisting of expert librarians and computer engineers embarked on a survey and review of research projects carried out throughout the world on digital library. This was done within a research project titled "digital National Memory"(HAMD) in research assistance, technology and programming. Later in 2008 the very project with a few changes under the title "digital national library," was put into operation. The project was intended to digitize all valuable and exclusive resources and to make available to the public. Lots of old resources including manuscripts, publications of Ghajar period, and old documents were scanned and digitized and were

organized in a software named “Nama”. Along with the increase of these digitized resources, some arrangements were to be made for the long-termed preservation and permanent access to them. This research seeks to put forward a practical and suitable paradigm for the preservation of the digital resources presently existing in the National Library and Archive of Republic of Iran, once it has surveyed the methods and strategies of digital preservation, digital repositories, standards, formats, tools storage, access levels, control access and information security in the national libraries, which are members of IIPC and National Archives with the membership of ICA applying an analytical and descriptive method and system design. The formulated questionnaires were e-mailed to the National Libraries in question.

Authors

Mitra Samiee obtained a BA degree in Library and Information Science from the librarianship Center of higher education, National Library of Iran in 1998, then received an MA in Library and Information Science from Azad University, Science and Research Branch, Tehran in 2002. She started her Ph.D. course in Library and Information Science at Azad University, Science and Research Branch, Tehran, Iran, in 2005. She is Fellow member of staff, National Library and Documents Organization of the Islamic Republic of Iran. She has more than 7 years of experience in electronic publication gained while working in the Organization of Documents & the National Library of Iran. She work in the doyenne of researches department, university lecturer and in charge of HARAM (i.e., National Digital Memory) project’s standards and digital content in the National Library and Documents Organization of the Islamic Republic of Iran .She is Manager website of The department of Iranian & Islamic study researches, the National Library and Documents Organization of the Islamic Republic of Iran. She teaches as a visiting lecture in the At the management faculty of Tehran university, National library’s department of administration, the organization of management and country planning, Azad university of Iran, the broadcasting Company of the Islamic republic of Iran .She is also a member of the Iranian Library and Information Science Association. She is written 24 papers and 2 books which are published in Iran. Her research interests include digital libraries, digital preservation, PREMIS and its usage in digital libraries, and data security and so on. Mitra Samiee is the corresponding author and can be contacted at: samiei.mitra66@gmail.com

Saeed Rezaei Sharifabadi is currently an Associate Professor at the Faculty of Education and Psychology, Alzahra University, Tehran, Iran. He obtained a BA degree in Library Science from Allameh Tabatabaee University, Iran in 1986, his MA in Library and Information Science from Tehran University in 1990, his Ph.D. in Library and Information Science at The University of New South Wales, Australia In 1996. His research interests include Information Technologies in Libraries, Internet use and user, digital records, digital archives, information seeking and communication behavior. He also supervised about 100 Master’s theses and Ph.D. dissertations and has written more then 45 papers in Persian and English, which are published in Iran and abroad. He teaches as an Associate Professor in the Alzahra University, Tarbiyat Modares University. Saeed Rezaei Sharifabadi is the corresponding author and can be contacted at: srezaei@alzahra.ac.ir



A Paradigm for the preservation of National digital memory of Iran

Mitra Samiee ; Fellow member of the National Library and Archive of the Islamic Republic of Iran, Tehran, Tehran/Iran
saeed rezaei Sharifabadi; Faculty of Education and Psychology, Alzahra University; Tehran, Tehran /Iran

Introduction

Similarly in Iran since 2001 some initiatives have been adopted by the NLAI to digitize the library resources and the development of a digital library has been officially proposed. To begin with a working group consisting of expert librarians and computer engineers embarked on a survey and review of research projects carried out throughout the world on digital library. This was done within a research project entitled "Digital National Memory" (HAMD) in research assistance, technology and programming. Later in 2008 the very project with a few changes under the title "digital national library" was put into operation. The project was intended to digitize all valuable and exclusive resources and to make them available to the public. Lots of old resources including manuscripts, publications of Ghajar period and old documents were scanned and digitized and were organized in a software named "Nama". Along with the increase of these digitized resources, some arrangements were to be made for the long-term preservation and permanent access to them. This research seeks to put forward a practical and suitable paradigm for the preservation of the digital resources presently existing in the NLAI, once it has surveyed the methods and strategies of digital preservation and the control access and information security in the national libraries which are members of IIPC.

History/background of the research

Titia Vander Werf-Darelaar (1999) in his article entitled "Long-term preservation of electronic publications" looks into NEDLIB research project and the technicalities of digital preservation strategies. In a report of CEDARS research project (2002), the project along with the issues regarding preservation and methods of access to resources as well as digital preservation strategies have been analyzed. Svein Arne Solbakk (2003) in his research article entitled "Critical technological and architectural choices for access and preservation in a digital library environment" describe some basic architectural choices for the access to and preservation of digital objects at the National Library of Norway.

Purpose: This research aiming at a survey of the existing status of digital preservation among the members of the IIPC was carried out by the identification of maintenance methods and strategies, digital repositories, standards, storage formats and tools and levels of access and security. Queries of research

Method/ approach: The research is based on a descriptive and analytic survey. According to this method, the community in question has been surveyed within two groups of digital archiving and permanent access as regarded digital preservation using questionnaires. The questionnaires were sent electronically to the community in question including 11 members of the IIPC.

Finds: %90/91 of national libraries apply on-line magnetic media and tape library, %81/82 SAN technology, %90/91 METS standard and the OAIS reference model, %90/91 PDF format, TIFF/EP3 and WAVA, %100 back-up supply, %45/45 access to the entire collection for free, access with restricted copyright and free access to the collection of digital records in part and %100 the mechanism of verification and access control management.

Conclusion: According to the results of this research, on-line magnetic media and tape library are appropriate for storage and long-term preservation of back-ups and digital resources at national libraries due to their high quality and considerable life span and the SAN for better integration, and allowing for the sharing of back-up facilities.

Table 2. required standards of digital preservation [5], [6]

Type Library	METS	OAIS	PREMIS	DOI	VRA	EAC	MOIS
Italy	1	1	1	0	1	1	0
Netherlands	1	1	1	1	1	1	1
Sweden	1	1	1	1	0	0	1
The British Library	1	1	1	1	0	0	1
Denmark	1	1	1	1	1	0	1
Sweden	1	1	0	0	0	0	1
France	1	1	1	0	1	1	1
Canada	1	1	1	1	1	1	0
Norway	1	1	1	1	1	1	1
Finland	1	1	1	1	1	1	1
The Library of Congress	1	1	1	1	1	1	1
Total	11	11	10	7	9	9	9
percent	100	100	90.91	72.73	81.82	81.82	81.82

Query2. which standards do these libraries apply for the long-term storage of their resources?

The finds show that almost %100 of the community in question apply METS standard and the OAIS reference model, %90.91, PREMIS, %81.82 MOIS and VRA, %72.73, DOI, and %45.45 EAC. The wide usage of these standards and metadata by the national libraries in question stands for the importance of these standards in digital preservation of resources.

Table 3. Types of digital preservation strategies

Type Library	Technology Emulation	Information Migration	Encapsulation	Digital Archiving	Data Backup	Reproduction
Italy	0	1	1	1	1	1
Australia	1	1	1	1	1	1
Sweden	0	1	0	0	1	0
The British Library	0	1	1	1	1	1
Denmark	0	1	1	0	1	1
Sweden	0	1	1	0	1	1
France	0	1	1	1	1	1
Canada	0	1	1	1	1	1
Norway	0	1	1	0	1	1
Finland	0	1	1	0	1	1
The Library of Congress	1	1	1	0	1	1
Total	1	10	9	4	12	10
percent	9.09	90.91	81.82	36.36	100	90.91

Query 3. Which strategies of digital preservation are applied at national libraries?

%100 of libraries apply the strategy supplying back-up, %90/91 of the libraries use information migration strategies and permanent identifiers to locate digital resources and for the long-term preservation of their digital resources, %81/82 employ information encapsulation strategy in OAIS reference model, %36/36 have attempted to use the strategy of technology preservation and digital archaeology and %18/18 apply technology emulation strategy.

Table 4. methods of accessing digital resources

Type Library	Free access to the entire collection	Access by copyright, copyright	Access to the library's digital resources, copyright	Free access to a part of the collection of digital resources	Access to the library's digital resources, copyright	Free access to a part of the collection of digital resources
Italy	1	1	1	1	1	1
Australia	1	1	1	1	1	1
Sweden	1	1	1	1	1	1
The British Library	1	1	1	1	1	1
Denmark	1	1	1	1	1	1
Sweden	1	1	1	1	1	1
France	1	1	1	1	1	1
Canada	1	1	1	1	1	1
Norway	1	1	1	1	1	1
Finland	1	1	1	1	1	1
The Library of Congress	1	1	1	1	1	1
Total	11	11	11	11	11	11
percent	100	100	100	100	100	100

Query4. How is access control carried out at national libraries?

Insights on the Digitization of Traditional Medicine Knowledge in Nigeria

Chinyere A. Otuonye, Tamunoibuomi F. Okajagu, Samuel O. Etatuvie, Emmanuel Orgah, Gift Eyemienbai, Luke Oyovwevotu, Ewoma Borgu, and Janet Ukoha

Author

Otuonye Adanma Chinyere, a graduate of Industrial Chemistry, is a Senior Research Officer with the Nigeria Natural Medicine Development Agency (NNMDA), a research institute of the Federal Ministry of Science and Technology (FMST), Nigeria. Since 2005, under the Traditional Knowledge Digital Library of Nigeria Project, she has worked with teams on the documentation and digitization of published research findings on Medicinal, Aromatic and Pesticidal Plants (MAPPs) and Traditional Medicine Knowledge and Practices (TMKP) of Nigeria. Ms Otuonye has served on several committees within and outside the NNMDA and is also a recipient of two awards.

Insights of the Digitization of Traditional Medicine Knowledge in Nigeria



Otuonye A. Chinyere, Okujagu Tamunoibuomi F., Etatuvie Samuel O., Orgah Emmanuel, Eyemienbai Gift, Oyovwevotu Luke, Borgu Ewoma, Ukoha Janet

Nigeria Natural Medicine Development Agency (NNMDA), Victoria Island, Lagos State, Nigeria

<p>Introduction</p> <p>The documentation and digitization of Traditional Medicine Knowledge and Practices (TMKP) in Nigeria has become inevitable as a result of lack of documentation, issue of access to information on Traditional Knowledge (TK), depleting natural resources, biopiracy and the need to research, develop, promote, protect and preserve the nation's heritage.</p> <p>There is also the need for a central reference point on published researches related to Traditional Medicine and Natural Products and their further development and promotion.</p> <p>Objectives of the Project</p> <ul style="list-style-type: none"> • Collate, document and digitize published research findings on Traditional Medicine and Medicinal Plants of Nigeria • Document and digitize data obtained from ethnomedicinal surveys of the six geopolitical zones in Nigeria • Document findings of the survey conducted among traditional medicine practitioners in Nigeria on their use of Medicinal, Aromatic and Pesticidal Plants (MAPPs) as food and medicine to address health challenges. • Document traditional healing systems and practices in Nigeria • Provide long-term articulated preservation, protection options and access strategies to documented data. 	<p>Method</p> <p>Oral accounts from traditional medicine practitioners (TMPs), data and digital images from ethno-botanical surveys and published research findings on health, Nigerian MAPPs and traditional medicine practices obtained from the field form the bulk of the digital library collections. Digital images were identified and labeled using binomial standard</p> <p>Challenges</p> <ul style="list-style-type: none"> • Literacy & Level of appreciation of Knowledge sharing • Intellectual property • Technology • New and different terrain • Technical Capacity and Capability • Funding (equipment and software, subscriptions, etc) • Time (Funds release, delay in project implementation timeline) • Not seen as a Political Priority Project • Unstable power supply <p>How Resolved</p> <ul style="list-style-type: none"> • Extensive education & training of Knowledge holders and practitioners • Drafted an IP regime for the protection of TK and Genetic Resources • Partnerships and support from national and international organizations • Participation in symposiums, trainings, workshops and conferences • Inverter/Alternate power source 	<p>Publications</p>        
<p>Acknowledgement</p> <p>This project was made possible by the Nigeria Natural Medicine Development Agency, the Federal Ministry of Science and Technology and the Federal Government of Nigeria.</p>	  <p>Inside view of the physical library containing the computers for internet users</p> <p>Library Portal (www.nignaturemed.net) Institutional Website (www.nnmda.gov.ng)</p>	

The Memory of the World in the Digital Age: Digitization and Preservation

VENUE: Vancouver, British Columbia, Canada
DATE: SEPTEMBER 26-28, 2012



Crowd-sourced digital preservation

An Iranian model

Nader Naghshineh¹ and Saeed Nezareh²

¹*University of Tehran, Nnaghsh@ut.ac.ir;* ²*University of Tehran*

Authors

Nader Naghshineh is a faculty member at the LIS Department, University of Tehran. He also serves as technical advisor to Central Library there. His professional track covers information services both domestically and internationally. In 2005, he helped establish an LIS lab for University of Tehran. In 2010, he established the Digital Observatory Lab followed by establishment of a Cognitive Informatics Lab. His current primary area of research could be best termed as 'Techgnosis', although most of his research has dealt with information technology customization and information transculturalism

Saeed Nezareh is a research assistant at the Digital Observatory Lab (ISL2). He also is a graduate student of University of Tehran and is enrolled in the ISL Research Excellence through Enhanced Learning Program. He has applied for a patent for a special medical diagnostic program.

Crowd-Sourced Digital Preservation: An Iranian Model



Nader Naghshineh, Director, Information Studies Lab, University of Tehran, Nnaghsh@ut.ac.ir
Saeed Nezareh, Research Assistant, Information Studies Lab, University of Tehran



Iran is a land of paradoxes. It has the fastest growing rate in scientific publications in the world, yet it languishes in developing the necessary support structure for sustainable innovation. As a country who claims to have a 3000 years civilization, its general public and even policy makers have little or no sensitization to preservation in all its form. The present study outlines the conceptual work carried out by Digital Observatory Lab (ISL-2) at the Central Library and Documentation Center at University of Tehran. ISL-2 is unique by the fact that it has been funded through private loans and lease-lends, which gives it flexibility and freedom in pursuing certain lines of research. In just two decades, the Iranian visual and audio Cloud had gone through four revisions. In broadcasting alone, the number of revisions is even more. However, more often than not, there are no extant conservation and preservation protocol in place. The past three decades seems to have made people more focused on Now than Tomorrow. In some cases, the digital asset is lost to degradation and dilapidation even beyond the reach of our digital forensic techniques, simply because the owner of the asset does not trust the government office that has the means to convert and preserve the asset.

To complicate matters further, the Iranian public does not tow the academia definition of digital preservation. There are groups that believe that efforts must be made to protect the information content (or Data pattern) of the digital object. There are those who believe that the digital object should be preserved in the way its information content was meant to be appreciated. And of course there are those who believe that while the information pattern must be preserved and protected, the very milieu that gave rise to it must be future-proofed; A rather tall order for a country beset by sanctions, deteriorating economy and patchy technological infrastructure. Thanks to the breakthroughs in technology, the threshold for an average Iranian to contribute a digital object has been pushed down in spite of the crazy exchange rate. This is a domain that is largely terra incognita. And while it is part of Iranian digiscape, the government bodies responsible for digital asset preservation have chosen to ignore it.

In 2011, the laboratory decided to initiate a study to help the owners of such digital assets to be provided with alternate service routes. Anonymity and affordability were the primary specifics. We hoped that in this way the digital content would be at least preserved after a fashion, until either the owner or custodian felt confident that it could release it to the local Digital Cloud.

Crowd Sourcing seems to be a promising venue based on our initial studies. Using the University of Tehran Community as a baseline, we came to the understanding that there are several social networks that could provide "Garage-Scale" digital preservation service, either in the area of obsolete equipment, coding or collection hunts. We tried several approaches to understand these networks and to see whether a Social Network could be induced that partakes the best attributes of each. Acknowledging the particulars of Iranian belief networks, we arrived at a model for a Crowd Sourcing Network that focuses on know-how trading. It also provides the veteran practitioners with an additional, non-government source for securing the necessary parts and labor for effecting digital conversion and preservation. The problem with this model that we are presently attempting to resolve is the trust network superpositions. It is interesting to note that under the current body of Iranian laws many preservation exercises could be threading a fine line between misdemeanors and crime. Underground Music, PODCASTs and Nickel-dime documentaries, family histories recounting an interpretation of some events are technically illegal. Yet they represent part of people artistic as well as narrative expression in the digiscape. They often provide interesting research materials if they are preserved.

It was proposed to establish a social-network focusing on providing grass-root support for micro-budget preservation inquiries. Such network would bring together the artists, the experts, the vendors and even the collectors together. It is like having *Technibes* crowdsourcing engine and E-bay and LinkedIn wrapped into one seamless service. The catch is how to ensure security, anonymity as well as prompt payment. The anonymity portion is proving to be quite a challenge.

Our model hinges on the fact that the digital asset would not leave the purview of the owner. In order to do so, it should be possible for the digital preserver to work on the asset without taking it out of its secured area. In a sense the digital asset should never be transferred in its totality and can only be accessed and worked on in its original form while within a virtual, digitally secure area. It is like taking your painting to a gallery everyday for restoration.

Data at Risk

The Duty to Find, Rescue, Preserve

Chris Muller

Author

Chris Muller founded Muller Media Conversions in 1978. The company is a leader in the field of data/media recovery and conversion. He has written many articles, published in Data Center Management and several other magazines. Mr. Muller is often called upon by law firms, private companies, universities and government to “make sense” of obscure or difficult data files. Examples include “WWW”—nope—not the web—but three of the cases whose data have been unscrambled and recovered by MMC: Watergate, Whitewater and WorldCom ☺. This work has naturally led to an interest in protecting information from harm and preserving it for the future.

DATA AT RISK: The Duty to Find, Rescue, Preserve

Brought to you by CODATA's Data at Risk Task Group (DARTG)

Please visit Session D at 1:30pm on September 26, 2012



"DARTG" is a Task Group of the Committee on Data for Science and Technology, a body of the International Council for Science.

See <http://ils.unc.edu/~janssp/dartg/>

We in DARTG define "data at risk" as scientific data which are not in a format that permits full electronic access to the information which they contain. Such data may be inherently non-digital (e.g. handwritten or photographic), or near-obsolete digital media or insufficiently described (lacking meta-data).

Recovering the Forgetting of the World

The first presentation is by Dr. R. Elizabeth Griffin of Dominion Astrophysical Observatory, who chairs the Task Group as well as other international groups. Her presentation will discuss DARTG's founding principles and major goals, e.g. creation of an inventory of data whose unique value is in danger of being lost to posterity.

You will hear many thought-provoking ideas, such as: "Delving back into the past is a route to instant science." ... "the complementarity of old and new" ... "Doing nothing had inflated expense, and ruined lives."

For more on this, please visit: **Session D – 1:30pm Wednesday**, where we will also describe initiatives taken by members the Task Group to preserve scientific data at risk.

Extending the Climate Record

1943 2004



Given by Stephen Del Greco, Chief of NOAA's National Climatic Data Center (NCDC) Climate Services and Monitoring Division (CSMD)

Join us at Session D and hear of the many NOAA data rescue projects and partnerships, including:

Climate Database Modernization Program (CDMP). Paleoclimatology efforts around the world.



Historical Marine Observations and much, much more....

Like Elizabeth, Stephen will stimulate your day with fascinating facts and ideas. Here's one of his phrases we specially enjoy: **"Old Weather: Our Weather's Past, the Climate's Future"**

Tide Gauge Data Rescue

Presented by Mr. Patrick C. Caldwell, Manager, Joint Archive for Sea Level at NOAA's National Oceanographic Data Center (NODC), this segment will convey such matters as how older records are of great value to understand more completely the time scales of sea level change.

Patrick is also a member of the Group of Experts (GE) member of the Global Sea Level Observing System (GLOSS). Topics include:



The Joint Archive for Sea Level (JASL), The Global Oceanographic Data Archaeology and Rescue (GODAR) project.

And several other initiatives. We'll also hear about the very interesting results of the **GLOSS 2012 data archaeology and rescue questionnaire**.

Maps: a fundamental source in the Memory of the World

The fourth presentation is by Tracey P. Lauriat, Postdoctoral Fellow, Geomatics and Cartographic Research Centre (GCRC), Carleton University and Dr. Fraser Taylor, Distinguished Research Professor of International Affairs, Geography and Environmental Studies at Carleton University, and Director of the Geomatics and Cartographic Research Center.



Expect to hear of many projects and insights.

One example: Born Digital Maps don't necessarily improve matters.... **"In Canada we have lost over 90 percent of the digital maps produced in the last quarter of the 20th century."**



Maps are ubiquitous in the 21st century, including cybercartographic atlases and many other manifestations. Nevertheless...

... **"It's easier to get maps from hundreds of years ago than maps from the last decade."**



Ancient & Digital

Once, these words seemed contradictory. Now, any digital media more than a decade old is "ancient." This information legacy is at great risk! Session chair, Chris Muller, founder of Muller Media Conversions (MMC) includes here a brief mention of an encouraging data rescue project:

Bangladesh Bureau of Statistics Data Recovery Center (BBSDRCC)

More than 6,000 reels of mainframe tape with important census and other data had been accumulated since the 1980's. Frequent power outages and other pressing needs made long-term preservation very difficult.



IPUMS.ORG via MMC provided the equipment, software and training to institute the project. New BBSDRCC management has recently achieved additional system and staff capabilities.

See www.ipums.org and www.mullermedia.com

Digital diplomacy

Providing access to cultural content, engaging audiences on a global scale

Natalia Grincheva

Concordia University

Author

Natalia Grincheva is a doctoral researcher in the Centre for Interdisciplinary Studies in Society and Culture, where she explores how museums can use digital media to preserve cultural heritage and develop intercultural dialogue. In summer 2011, Natalia was on the UNESCO Secretariat team to assist with the activities of the International Fund for Cultural Diversity. In October 2011, Natalia joined the team of the Coalition for Cultural Diversity in Canada, where she has been serving as an Associate Researcher.



Digital diplomacy:

Providing access to cultural content Engaging audiences on a global scale

INTRODUCTION

Information technologies has transformed our society operating in a world that is increasingly global and virtual. Despite of the fact that many people share a virtual space of the Internet and connect to digital resources on a daily basis, the issues of cross-cultural understanding, tolerance, and respect still remain imperative. Cultural diplomacy has traditionally been an instrument of interacting with the outside world. It is based on the exchange of ideas, information, art and other aspects of culture among nations and their peoples to foster mutual understanding.



OBJECTIVES

The research aims to evaluate the impact of digital diplomacy implemented within a museum context.

METHODS

The research draws on case study analyses of online projects developed by museums in four different countries: the UK, the United States, Australia, and Canada.

VIRTUAL ETHNOGRAPHY

SOCIAL NETWORK ANALYSIS

RHETORICAL FRAMES ANALYSIS

DESIGN, STRUCTURAL ANALYSIS

FINDINGS



INTERNATIONAL LEVEL
WORLD DIGITAL LIBRARY,
UNESCO

- Promoting cultural international agenda
- Strengthening international outreach
- Narrowing digital divide among and between countries



MULTINATIONAL LEVEL
EUROPEAN COMMISSION

- Constructing European identity
- Promoting EU cultural values
- Strengthening regional ties and cultural cohesion



NATIONAL LEVEL
A HISTORY OF THE WORLD
PROJECT, BBC AND
THE BRITISH MUSEUM

- Promoting British culture and English language
- Advertising national expertise and excellence
- Expanding national outreach to foreign communities

FOR MORE INFORMATION
Contact Natalia Grincheva, 2014 PhD Candidate
CISSC, Humanities Program, Concordia University
E-mail: grincheva@gmail.com, Tel: 323 336 4496

Preserving digital research data

Rusnah Johare

Author

Dr. Rushnah Johare, a senior lecturer of records and archives management, has many years of professional and academic experience in the information management field, ranging from job as an archivist at the National Archives of Malaysia to her current position as the Deputy Dean (Quality and Research) at the Faculty of Information Management, Universiti Teknologi MARA Malaysia (UiTM). She received a B.A (Hons.) from the University Science of Malaysia, a Professional Certificate in Archives Administration and an M.A in Overseas Records Management, both from the University College London, and a Ph.D. in Electronic Records Management from the Northumbria University, Newcastle, England.

Official InterPARES Website: <http://www.interpares.org/>

Empowering local communities in Campinas in Digital World

Claudia M. Wanderley and Tel Amiel

Authors

Dr. Claudia Wanderley is a researcher at the Center for Logic, Epistemology and the History of Science at the University of Campinas (UNICAMP, Brazil). She created a network for research development on Multilingualism in Digital World in Portuguese speaking countries. The focus of the network is local languages revitalization, cultural heritage and patrimony preservation. (@Multilinguismo).

Dr. Tel Amiel is a researcher at the University of Campinas (UNICAMP, Brazil). He coordinates the Open Education portal (www.educacoaberta.org), a collective that promotes research, development, symposia and workshops on open education and open educational resources with a focus on K-12 schooling and non-formal learning environments.



Empowering local communities in Campinas in Digital World

Claudia Wanderley & Tel Amiel

Center for Logic, Epistemology and the History of Science (CLE/UNICAMP) and Núcleo de Informática Aplicada à Educação (NIED-UNICAMP)
www.unicamp.br ~ ctwanderley@gmail.com ~ telamiel@telamiel.info



Introduction

The region of São Paulo started to figure on a map of what would become Brazil with expeditions for precious stones and for gold, then with enslaving expeditions [to capture indigenous peoples and to sell them as slaves], and later to establish coffee plantations. Campinas, now the second largest city in the state of São Paulo with over 1 million inhabitants, was known as a major slave trade center.

The struggle of fugitives against captivity founded quilombos which both in Kimbundu (quilombos) and in Portuguese designates "settlements". These settlements are still exist within the urban confines of the city of Campinas.

In 2003 the Brazilian Government issued a law determining that the history and the culture of indigenous and african people should be taught in Brazilian schools. This initiative promotes the voices of these peoples in the narrative of Brazilian history, provides recognition, and helps preserve their traditions and their language as national patrimony.

This project is about empowering two urban quilombos in the digital world. The aim is to assist them in recognizing and organizing what they consider to be worth preserving, digitally. This process occurs through two steps. First through the creation of a digital collection in a local digital library; and second, by making these resources openly available to stakeholders in public schools across Brazil.

Conclusions

The biggest challenge of this project is to level the academic notion of knowledge with the notion of knowledge of the communities involved. To work in partnership with quilombos towards the digital inclusion of their culture, promoted by members of their community (and not exclusively by scholars), permits to envisage differences and contradictions between the disciplinary academic formation and the native tradition. The inscription of aspects of local culture in a public database, or in an open digital library, provides visibility to the traditionally unauthorised representation of local cultures in Brazil and the difficulty local groups have to represent local knowledge in traditional schemes (for example by creating "metadatas").

This process of digital preservation assigns a new role to digital patrimony and digital preservation, a territory to be occupied by historically un-inscribed (not-represented) people. To acknowledge this situation means an understanding the demand for the inclusion and representation of local cultures (through their own actions) in the digital world, and not through "extraction and appropriation" processes of local culture in the name of "general" global digital access to otherwise voiceless peoples. In this sense, we aim to contribute to the long term preservation of digital patrimony in developing countries, and see the support and oversight of multinational agencies such as UNESCO as essential in promoting such practices.

Bibliography

AMEL, T.; REEVES, T. C. Design-Based Research and Educational Technology: Retexting Technology and the Research Agenda. *Journal of Educational Technology and Society*, v. 11, n. 4, p. 28-40, 2008. Disponível em: < http://www.jett.edu/journal/v11_n4/p28-40.pdf >
TOZZI, V. Redes Federadas eeventuais conectadas: Arquitetura e protótipo para a Rede Mocambos. Orientador Rocco de Nicola. In: *Universidade Digital Studio 6 Fierze*, Curso de Ciência da Computação, defesa 2012. TCC ano 2010-2011. <http://www.mocambos.org/Arquivos/Muro_RP/EC_RedemMocambos_VT.pdf>, <http://www.mocambos.org/Arquivos/Muro_RP/EC_RedemMocambos_VT.pdf>
WANDERLEY, C.M. Biblioteca Digital Multilíngue: uma proposta inclusiva. In: *Greenstone: un software libre de código abierto para la construcción de bibliotecas digitales. Experiencias América Latina e Caribe*. Ed. UNESCO Montevideo 2010. isbn 978-92-809-149-8. <<http://www.unesco.org/ur/ol/lebrun/comunicacion/informacion/Greenstone-parametros.pdf>>

References: www.mocambos.org, www.greenstone.org, www.iphan.gov.br, www.ethnologue.org
Map 1: Quilombos connected online in Brazil. Map 2: languages in Brazil to be included and preserved in digital world.

Objectives

The aim of this project is to foster an educational center for local communities promoting low cost, resource-sharing arrangements, and the dissemination of research results and best practices to democratize access to digital preservation techniques and their outcomes (resources and collections).

Increase local self-awareness by recognizing that their culture consists of unique resources of human knowledge and expression;

Promote local digital inclusion of local actors through the creation, selection, treatment, and sharing of digital collections and resources;

Identify the barriers and limitation of implementing such a project in local communities, making this knowledge available to other actors interested in safeguarding local heritage and knowledge;

Methods

Contact local communities and present the concept of a digital library. Begin data mining by identifying the main interests of the community and map interests that might become material or immaterial patrimony;

Structure pilot digital collections on topics chosen by the community and initiate the digitization of local contents to integrate pilot collections;

Contact regional, national and international institutions for national and international recognition of the patrimony;

Create, within the community, an education and training programme focused on local patrimony, low cost resource-sharing arrangements, and dissemination of local content aiming at local autonomy;

Take necessary steps for digital preservation in multilingual basis

Materials

Digital cameras for video and photo capture;

Pendrives to gather and organize data;

Local or remote server to publish digital library and its collections.



Results

We have finished three pilot digital collections based on topics of local interest and expertise:

- 1) Baobab three
- 2) Drums art and craft
- 3) House of culture of the community (Quilombo Urbano Casa de Cultura Tainá).

Subsequently four other quilombos demonstrated interest in the project and were invited to participate. The project in its full form was also submitted to an open call for projects by IPHAN (National Institute of Artistic and Historical Patrimony; *Instituto do Patrimônio Histórico e Artístico Nacional*) and was approved (funding pending).

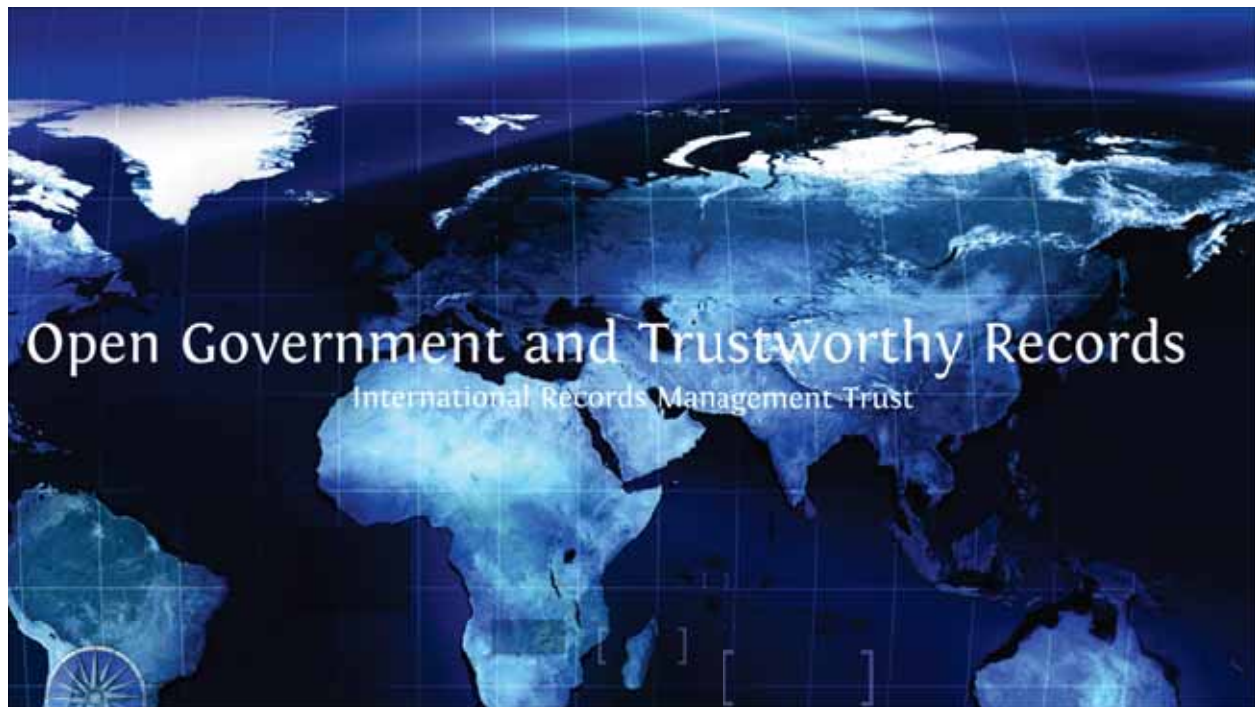
Open government and trustworthy records¹

Anne Thurston

Author

Anne Thurston has pioneered approaches to sharing solutions for managing public sector records with developing nations. Between 1970 and 1980 she lived in Kenya where she conducted research before joining the staff of the Kenya National Archives. In 1980 she became a Lecturer, later a Reader in International Records Studies at the School of Library, Archive and Information Studies, University College London. She established the International Records Management Trust in 1989 and continues to be its Director. In 1996 she left University College London to concentrate fully on the work of the Trust. In the 1990s, recognizing the impact of the rapid changes in the use of information technology on the management of public sector records, she structured the Trust to address the impact of those changes. Dr. Thurston was a member of the UK Lord Chancellor's Advisory Council on Public Records from 1994 to 2000. She was awarded an OBE for services to public administration in Africa in 2000 and a Lifetime Achievement Award by the Records Management Society of the UK in 2007. She was awarded the Emmett Leahy Award for Outstanding Contributions to the Information and Records Management Profession in 2007.

¹ The following presentation is available online at: <http://irmt.org/open-government-trustworthy-records-presentation>.

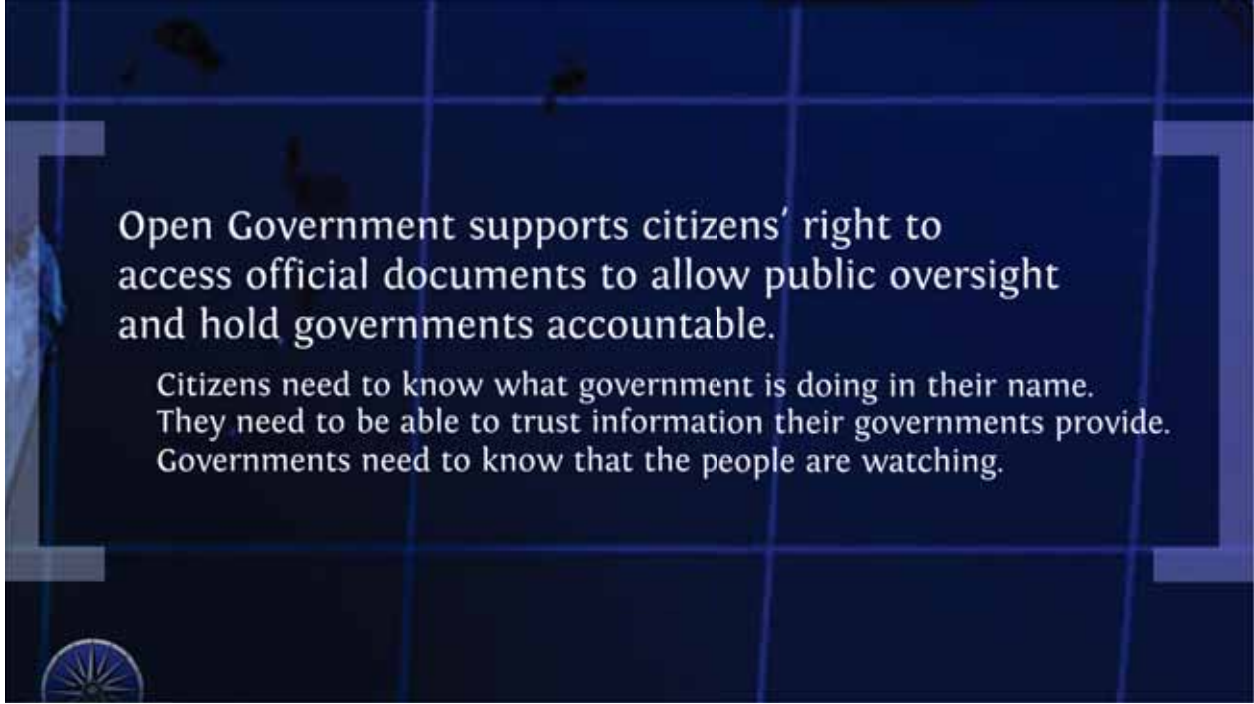


Trustworthy records are essential to:

- promoting transparency
- empowering citizens
- reducing corruption
- improving governance through new technologies.

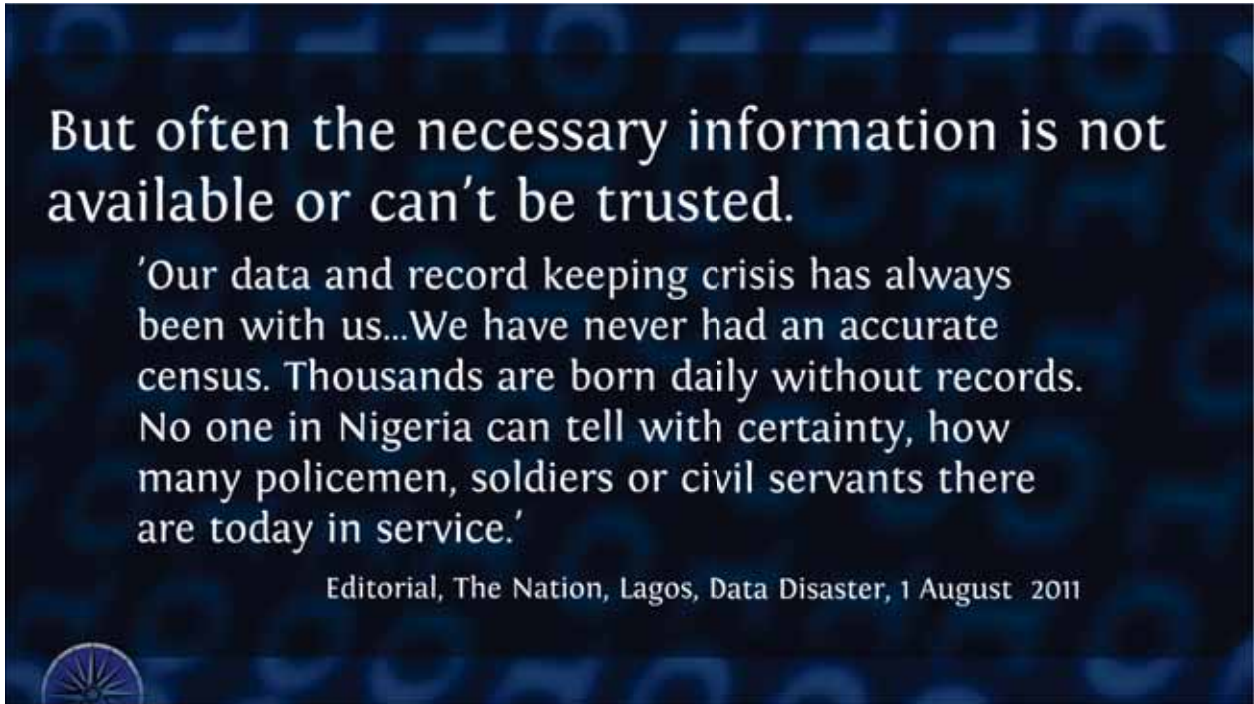
Without systematic controls, records can be easily manipulated, deleted, fragmented or lost.





Open Government supports citizens' right to access official documents to allow public oversight and hold governments accountable.

Citizens need to know what government is doing in their name.
They need to be able to trust information their governments provide.
Governments need to know that the people are watching.



But often the necessary information is not available or can't be trusted.

'Our data and record keeping crisis has always been with us...We have never had an accurate census. Thousands are born daily without records. No one in Nigeria can tell with certainty, how many policemen, soldiers or civil servants there are today in service.'

Editorial, The Nation, Lagos, Data Disaster, 1 August 2011

Joseph Rugumyamheto

Former Permanent Secretary, Government of Tanzania



John McDonald

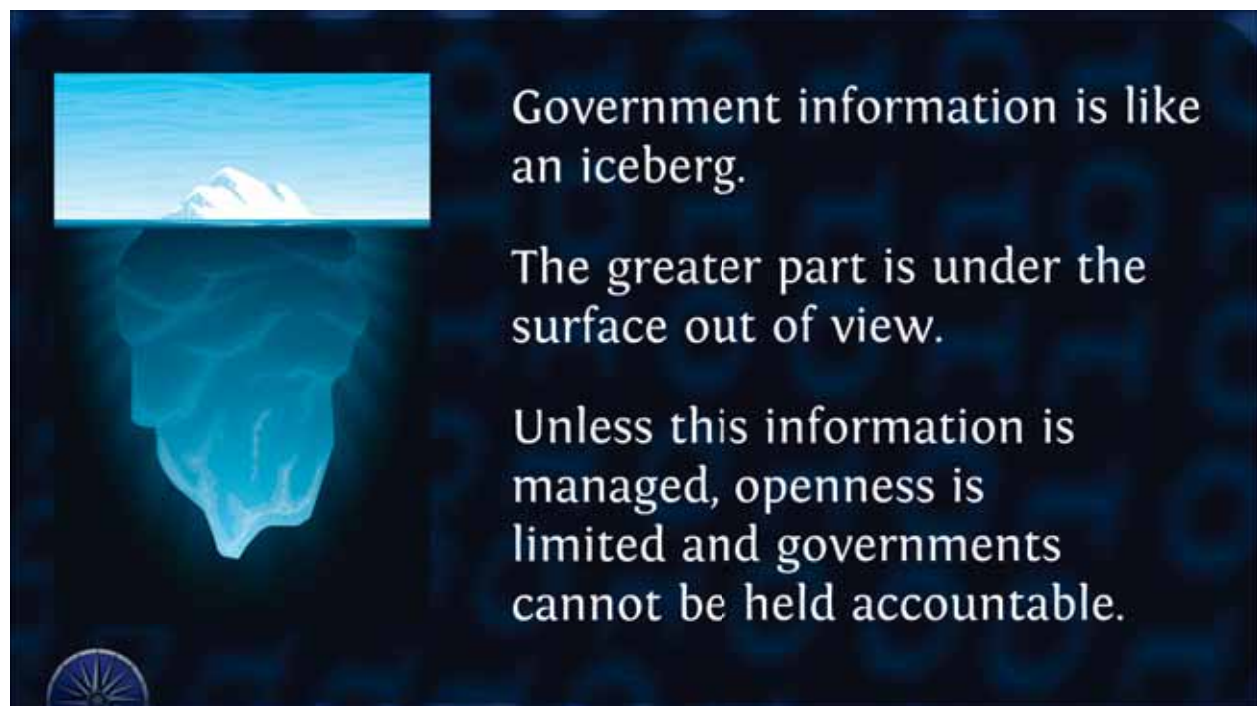
Former Director, Electronic Records Programme
Government of Canada

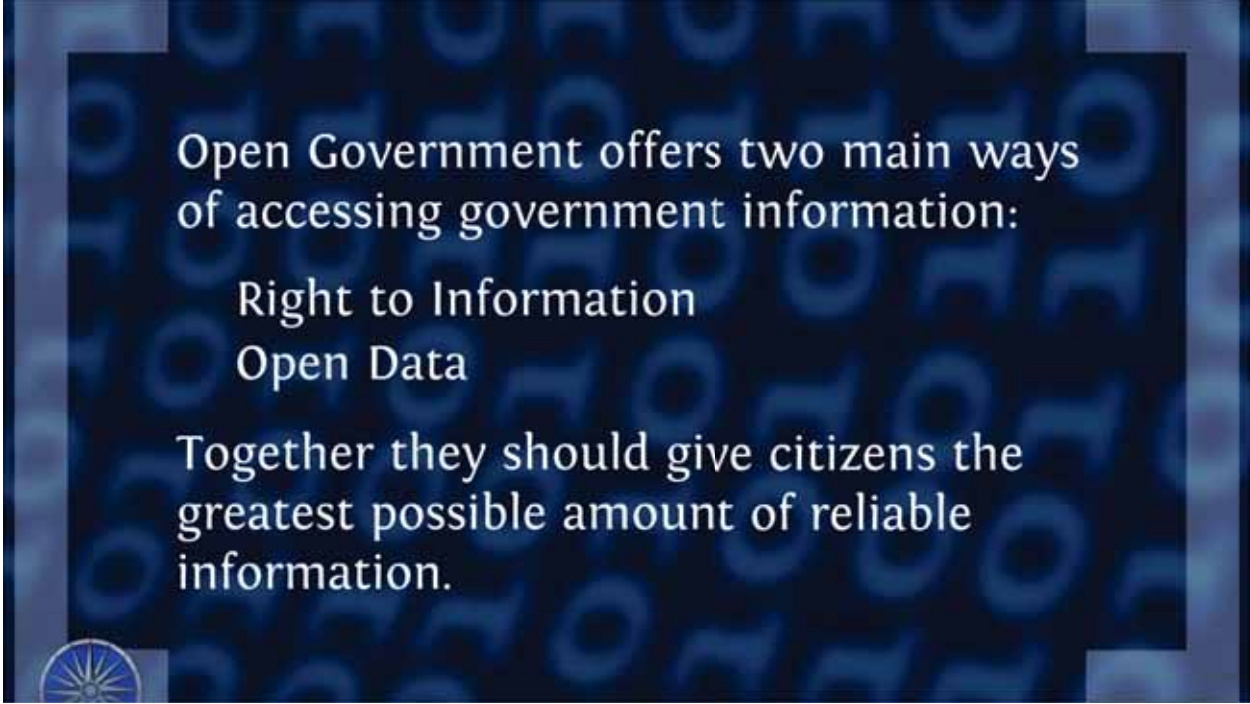


Please visit the online presentation, at <http://irmt.org/open-government-trustworthy-records-presentation>, to listen to the audio presentations for the above two slides.



Please visit the online presentation, at <http://irmt.org/open-government-trustworthy-records-presentation>, to listen to the audio presentations for the above slide.





Open Government offers two main ways
of accessing government information:

Right to Information

Open Data

Together they should give citizens the
greatest possible amount of reliable
information.



Right to Information / Freedom of Information

Governments give citizens a right of access to
information held by public authorities on request,
subject to certain conditions and exemptions.






The right is of little use if reliable records are not created in the first place or cannot be found when needed.

Open Data

Governments use the Internet to proactively disclose data without restriction, for reuse free of charge to:

- catalyse openness and transparency
- support research
- foster economic growth and global competitiveness.



Records and Data are closely related.


Records document functions, activities, processes (land rights, court decisions, financial processes) as evidence of:

- how policies or decisions were developed or executed
- what the government actually did
- what transaction were carried out.



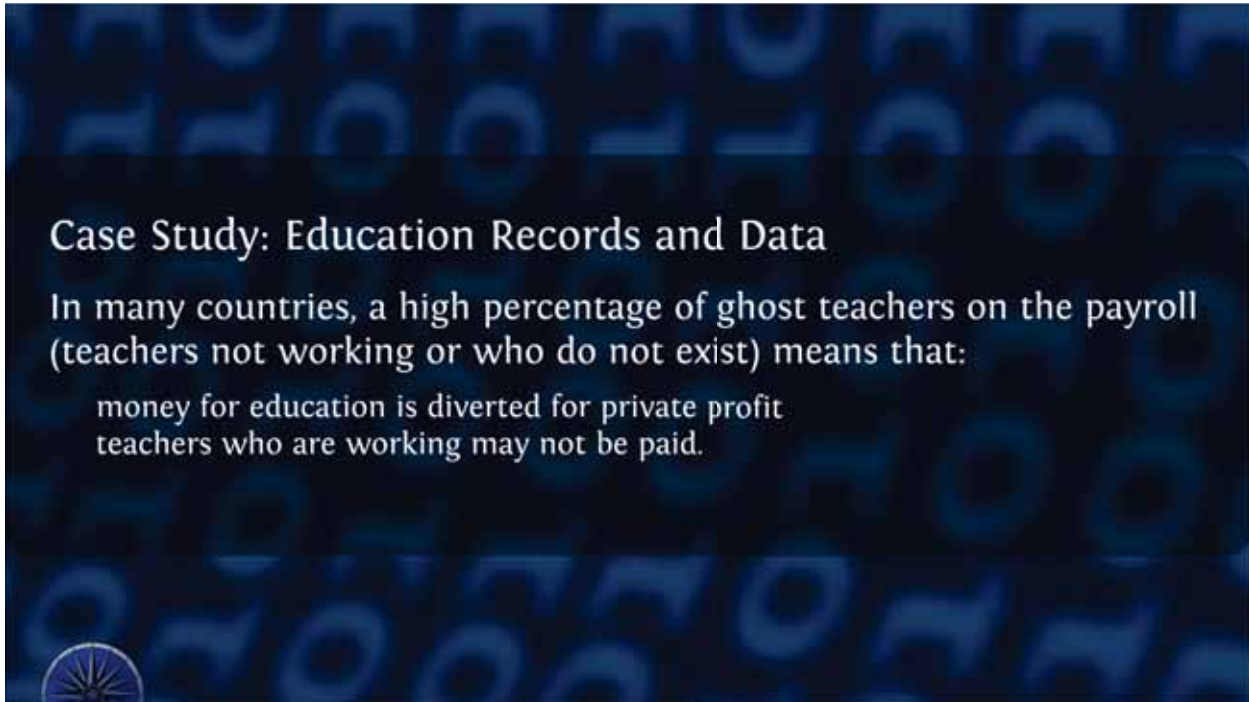
Data are generally derived from records:

- agricultural statistics from land survey records
- healthcare data from patients' records
- employment data from human resource records.



The quality of the data depends
on the quality of the records.

Data should be traceable to reliable
sources (paper and digital) or it risks
being misleading and open to political
manipulation.



Case Study: Education Records and Data

In many countries, a high percentage of ghost teachers on the payroll
(teachers not working or who do not exist) means that:

- money for education is diverted for private profit
- teachers who are working may not be paid.

Complete and accurate teachers' employment records are essential for:

validating the payroll
identifying teachers due to retire.



Complete and accurate teachers' employment records are essential for:

validating the payroll
identifying teachers due to retire.




Education planning and management
require accurate records of:

numbers of teachers
numbers and locations of schools
numbers and locations of pupils.



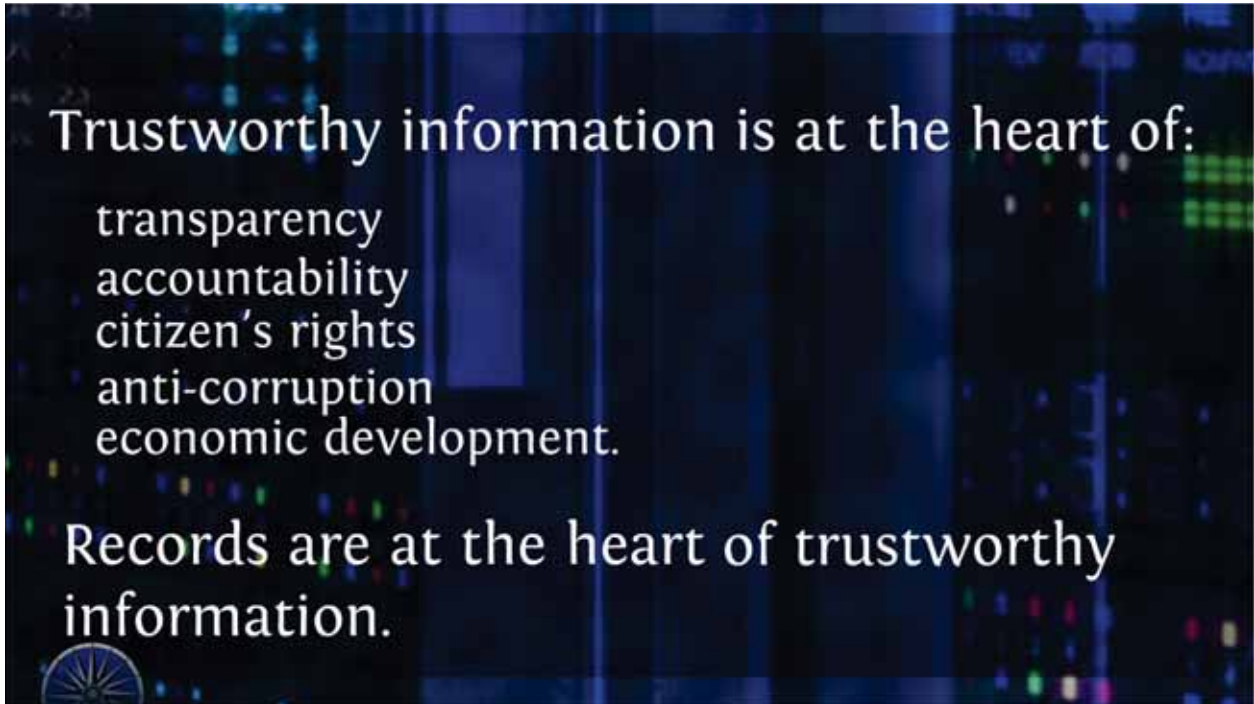
Data is a powerful tool
when accurate records
are mapped against
locations and the
payroll.





The records profession has defined standards and good practices. National archives should lead in defining government-wide policies for:

- ensuring that records are complete
- capturing context of how records were created and stored
- preserving and accessing records.



Trustworthy information is at the heart of:

- transparency
- accountability
- citizen's rights
- anti-corruption
- economic development.

Records are at the heart of trustworthy information.



Canada – Aural Memories

A Case Study of Soundscape Archives

Jan Marontate, David Murphy, Megan Robertson, Nathan Clarkson and Maggie Chao

School of Communication, Simon Fraser University

Authors

Jan Marontate, Hon. B.A. (York U.), M.Sc.(Demography, U. Montréal) and Ph.D. (U. Montréal) held a Canada Research Chair in Technology and Culture at Acadia University before joining the School of Communication faculty at Simon Fraser University in 2006. Her current research focuses on interdisciplinary networks of collaboration devoted to the preservation of time-based media in artistic and cultural heritage preservation initiatives.

Maggie Chao recently completed her BA with Honours in Communication at Simon Fraser University and is presently a Research Assistant in the SFU Visual Studies Lab. She is particularly interested in the study of urban culture and politics and its links to globalization and the rise of the creative city. She intends to commence graduate studies in the fall of 2013.

Nathan Clarkson, BA (Contemporary Arts, SFU) is a graduate student in the MA program of the School of Communication at Simon Fraser University. Nathan has been studying music, audio and other forms of media production since 2008. He has worked projects related to the preservation and management of the World Soundscape Project archives and other materials in the Sonic Research Studio at Simon Fraser University. His graduate research focuses on how technologies influence our understanding of natural phenomena. His interest is in the limitations of how we construct representations within given formats and the challenges faced by practitioners, and audience members, when relating to depictions of place.

David C. Murphy, BFA (Contemporary Arts, SFU), MA (Communication, SFU) is Senior Lecturer in the School of Communication. His research interests include media analysis, visual methodologies and risk communication. He is past-president of Vancouver New Music.

Megan Robertson, B.A. (English and Anthropology, UBC) and MA (English, UBC) is a doctoral candidate (ABD) in the School of Communication at Simon Fraser University. Her research interests include the study of collective memory and identity, vernacular photography and contemporary art.

Listening with Technology

New technologies for creating, recording and disseminating sound, visual images and texts are associated with dramatic transformations in the lived experience of acoustic environments and ways of documenting them.

The Listening with Technology: Environments in the Study of Sonic Transformations (LWT) research program investigates strategies for the preservation of documentation about environmental sound focusing on a case study of the **World Soundscape Project (WSP)**.

Research activities include analysis of the existing WSP collection, digitization of archival materials, interviews with past and current members of the WSP team, and the creation of the third series of field recordings of the Vancouver soundscape.

The WSP archives holdings include audio recordings in varied formats, visual images and texts (descriptions of sounds in literature, field notes, lectures and publications by participants in the WSP and a subject index).

The collection is an example of a cultural heritage collection in "time-based" media—a term used by cultural heritage professionals to refer to works created using ephemeral materials, performance-based practices or technologies that rapidly become obsolete.

Time-based media works cannot be preserved unchanged or they will be lost. They must be migrated, emulated, re-fabricated or re-performed.

Since preservation changes the materials in the collection & their relationships it is important to study what the collection represents and the multiple registers of meanings of the archives in historical, contemporary and future contexts.

Members of the **LWT** research team:
Principal Investigator: Jan Marontate
Co-Investigators: Barry Truax, David Murphy
Research Assistants: Megan Robertson, Nathan Clarkson, Maggie Chao
Recordist: Vincent Andriani

Aural Memories: A Case Study of Soundscape Archives

Poster Authors: Jan Marontate, David Murphy
Megan Robertson, Nathan Clarkson, Maggie Chao

School of Communication, Simon Fraser University, Burnaby, BC, Canada
Contact: Jan Marontate (jmarontate@sfu.ca)

Changing Technologies

Digital technologies expand possibilities but change ways of listening. Recorders grew smaller, more versatile & cheaper.



Nagra analog audio recorder. Portable but obtrusive technology (6 kg) used to make the first series of recordings in the 1970s.



DAT Tape (Digital Audio Tape). Used during the second set of recordings in the early 1990s.



Marantz MPO 681 handheld solid state digital recorder. Used for the third series of recordings (2010-2011). An iPhone was used to document GPS information and take photographs of field recording sites.

Field Recording Practices

Recording practices have shaped & been shaped by the interplay of technology and changing values of creators of archival materials.



Bruce Davis & Peter Huse creating field recordings with the Nagra (c. 1973)



Peter Huse creating a recording in a shooting gallery (c. 1974)



Vincent Andriani & Nathan Clarkson creating recordings using the Marantz (c. 2010)

Diverse Values & Interests

Creators had different interests that changed over time. When materials have multiple meanings which interpretations should prevail?



Members of the original WSP group at Simon Fraser University (L-R): R.M. Schaffer, Bruce Davis, Peter Huse, Barry Truax, Howard Broomfield (c. 1973)



The "Acoustic Crew" in 2010 (L-R): Nathan Clarkson, Jennifer Schine, Vincent Andriani, Milena Droumeva, Dave Murphy

Case Study: The World Soundscape Project

In the 1960s R. Murray Schafer, a composer and theorist, launched an ambitious program of research on acoustic environments that drew on interdisciplinary methods to document phenomena in everyday life, and re-imagine the future of soundscapes through acoustic design.

The World Soundscape Project, grew out of concerns for noise pollution but rapidly evolved into a large-scale documentary project about disappearing sonic memories of the world and, subsequently, recordings in the collection have served as an artistic resource for composers in their own creative work.

The WSP team developed new concepts such as the notions of soundscapes and soundmarks—aural analogies to landscapes and landmarks. Contemporary acoustic researchers and scholars of acoustic ecology emphasize the importance of listening and studying the way acoustic environments are understood by individuals and communities.

Barry Truax, director of the Sonic Research Laboratory at Simon Fraser University, has led WSP activities since 1975 and guided the maintenance and conversion of the archives to electronic formats.

Acknowledgements

The Listening with Technology study thanks the Social Sciences and Humanities Research Council of Canada and the Office of the Vice President and the School of Communication at Simon Fraser University for their support.

SSHRC CRSH



Creating Social Memories of Major Events in China

A Case study of the 5•12 Wenchuan Earthquake Digital Archive

Na Cai, Leye Yao and Liu Liu

Abstract:

This case study offers insight into practical approaches to creating social memories of the major events via constructing the 5•12 Wenchuan Earthquake Digital Archive (WEDA) in China, which provides detailed strategy regarding the construction of digital documentation in China, and emphasizes the importance of integrating the actual situation and background of the region and the nation. It is hoped that the valuable experience in WEDA construction provides a good example of creating social memory of major event, and will be of a help to those developing countries in similar circumstances of struggling for the preservation of digital documentary heritage.

Authors

Na Cai, Ph.D. of Archival Studies, is a lecturer in the Archives Department, School of Public Administration, Sichuan University, China, where she directs the National Fund of Social Sciences program titled “The management mechanism of major event archive and its exploration mode of resource utilization.” She was librarian of Sichuan University Libraries from 2006 to 2011. Her research interests and writings are in the area of digital archives and web information preservation. She has published about 30 papers in academic journals in China and abroad, and also is co-author of a book titled *Construction Strategy of Digital Archives*. Her contact information is: School of Public Administration, Sichuan University, No.29 Wangjiang Road, Chengdu, China. Tel: 86-28-138-8078-5072, Fax: 86-28-854-130-06, E-mail: caina@scu.edu.cn.

Leye Yao, Ph.D. of history, is professor and doctoral tutor in the School of Public Administration, Sichuan University. He also serves as Executive Director of China’s Development Research & Consultation Institution, SCU, Director of the office for Social Sciences, SCU, and as Vice-Chairman of the University Library and Information Service Steering Committee, Ministry of Education of China. He is the former Director of Sichuan University Libraries.

Liu Liu is a librarian with the Sichuan University Libraries.

1. Introduction

The documentation of major events is of great significance in creating related collective memories of a region, a nation, and even the world. In the digital age, the ways of information creation, transmission and storage of major events have change a lot, and have digital and network features, which distinguish them from what came before.

Digital preservation could be a new strategy of utmost importance for creating social memories in the digital age. It is extremely urgent to do so because some digital documentation of major events is disappearing irreversibly.

The case of the 5•12 Wenchuan Earthquake Digital Archive (WEDA) in China showed the awareness of preserving digital heritage, the process of overcoming obstacles from different aspects. Since it reflects the latest progress of digital preservation in China, it may provide a good example of preserving digital documentation of major events for other developing countries.

With narrative of related theory about major event documentation and social memory, and analysis of the features of major event documentation in digital era, this article will review this case to explore the construction of the WEDA, including the necessities, challenges, and solutions. In addition, the study will also assess the result and give future prospect of it, which will provide important information to others facing similar situations.

2. Theory of creating social memory in major event documentation

2.1 Social memory and major events

It is known that “social memory” is a term frequently used by sociologist or anthropologist when discussing how to keep and spread collective memory. Now it becomes a concept which is often used in sociology, history, ethnology, cultural studies, anthropology, and etc. It was firstly evolved by French sociologist Maurice Halbwachs, on the basis of the concept “collective memory”, later promoted by Paul Connertont, Barry Schwartz and other sociologists as modern theory of social memory. It focuses on three problems:

1. What determines specific social events being “chosen” or being “forgotten”? Why are these events rather than other events to be memorized? It refers to an event’s own characteristics.
2. How is the past constructed by a society? It refers to the issue of social memory’s social motive power base.
3. How is a social memory inherited? It refers to social memory building mechanism.

Scholars believe that a construction of social memory is a link of the past and the present.

2.2 History and major event material

According to the empirical historicism theory of a German historian named Leopold von Ranke, history should truly reproduce the past, and if an individual want to get a comprehensive understanding of history, the best way is to review the occurrence and process of history events. Therefore he attached a great importance to historical events. “uniqueness” and “development (evolution)” are the foremost concept in this theory. The former refers that each historical phenomenon, event and individual has its specific meaning. The later refers events connection, which reflects the continuity of historical process. Historical events could be completely understood only through digging useful historical materials, tracing event origins and abundant facts provided, and thought about human history could begin. Today this theory origin from history area is still producing great influence in other discipline.

2.3 History, social memory and major event documentation

French sociologist Michel Foucault put forward a point in his book named *Archaeology of Knowledge* that history is a proof of collective memory with thousands of years. And the memory depends on material literatures to regain its freshness of the past.

According to this view, the past is formed in the process of paying close attention to the present, therefore the concept “social memory” is distinguished from the concept “history” in tradition history, which inherits history and forms the past in the social applications as a storage of human behavior and

events. As a result, historians could separate himself/herself from the research objects in the research of past through documenting, classifying and analysing human behaviors and events, whether in the process of recalling life details of individuals, or constructing history memory of wider social vicissitudes.

The understanding and remembrance of major event is one of the most important collective memory in the social development. As solidified form of information and knowledge reflecting important development trace of human society, major event documentation becomes the main information source of major event, shaping the collective memory of related group. As a result, major event documentation plays a unique role in constructing the social memory than other forms, which connects past, present and future by inheriting and continuing the collective memory of family, organization, region, nation and even the whole world.

2.4 Major event documentation in the digital age

As a multimedia platform integrating words, pictures, voice and video, network becomes one of the most important information source about major event happens in the 21st century for its advantages of fast speed and wide coverage. Some important information about major event is created only in digital form. Taking the 5•12 Wenchuan Earthquake, which happened in China in 2008, as an example, the first message about this earthquake was transmitted online by instant message media, earlier than official media, including TV and other media. And in the rescue period after the earthquake, news website, BBS, web log of official department and survivors, and other network media are most active in revealing updated information, some information even directly promoted the relief work after earthquake.

However, as the documentation about major event maybe created in digital form, transmitted and stored online, it may fade as time goes by, if there is no measures taken intentionally to preserve the digital documentary heritage, and social memory about these major events will not complete and may not truly reflect what actually happens today. Especially in developing countries like China, under the environment of global network development, a great amount of network information mainly in Chinese is created and transmitted in intranet and extranet as economic and social development.

While documentation of major event today have various sources and forms, it can be foreseen that future social memory about major event happens today should reflect human life in the digital age of aspect different from ever before. Network becomes one of the most important media, carrying individual, organizational and community memories, including collective memory about major event as natural and social shifts and other changes, such as 2008 Beijing Olympic Games and 5•12 Wenchuan Earthquake, which have great influence in national wide even in the whole world as major events.

3. Digital preservation and WEDA

3.1 Obstacles of digital preservation in China

In view of the significance of major event documentation in constructing social memory of digital age, many institutions in China began to pay attention to digital preservation of major events documentation, which have various obstacles in this country: lacks of policy support and legislation protection, confines of administrative system and organizational restriction, shortages of funds and manpower, problems of intellectual property and related rights, disunity of metadata standards, etc. Apart from the common obstacles most countries have to overcome, China has to deal with the problems of its complex

administrative system and obstructive capital source. Furthermore, lack of practical experience of preserving digital documentation in Chinese is an obvious disadvantage compared with that in English-speaking countries.

3.2 Origin of WEDA

Since the 5•12 Wenchuan Earthquake happened in Sichuan Province of China in 2008, a great amount of related documentation has been created by various government offices, organizations and individuals, the content of which covering the basic information disaster-stricken area, the data about casualties and destroy of buildings, rescue record, reconstruction information, etc.

As this earthquake was the greatest earthquake happened after the founding of the People's Republic of China and had great influence in this area and the whole country, to preserve this district history and related social memory become a necessary task.

In order to construct complete social memory, which can truly reflecting this major event from various aspects, to collect related documentary materials together and construct a database named 5•12 Wenchuan Earthquake Digital Archive (WEDA) can be a good solution. It can also be convenient for users in digital age through offering access for users in digital form and by network.

4. Challenges of WEDA construction and solution

4.1 Challenges of WEDA construction

WEDA construction faces similar situation as other digital preservation program in China, which was particularly present as these problems:

Firstly, since there was no government agency such as government information office or archive administrative agency authorize any institution such as library or archives to construct such database, and there was also no regulations for the management of major event documentation, so WEDA construction have no policy and legislation support.

Secondly, as the documentation about this event included different forms of materials, including books, magazines, records, archives, pictures, videos and web information, which had various information resources, scared in government agencies, institutions, civil organizations and individuals, many of them had different administrative systems and ownerships, some of these materials not allowed to open to the public for involving privacy or secret of organizations, which caused the difficulties of collection.

Thirdly, for the lack of lack of policy and legislation, and private donation is not a common source in China, so the short of fund and capital investment for WEDA is also a challenge too.

While, there are other challenges. As this database aims to preserve related social memory through collecting truly reflect this major event from different aspects, not only professional earthquake materials, but also other related information such as web information mainly in Chinese, which have no previous example for reference.

4.2 Solution of policy support

In present China, there is no particular legislation or any authority by government department on the digital preservation of major event. Some library or civil organization had tried on digital preservation program, but mostly aimed at preserving web information, not preserving social memory of particular major event in digital form. For example, National Library of China had a experimental program named “Web Information Collection and Preservation,” which took “Event of China” as a part. Some archives or library aimed at traditional preservation of major event, such as “Tangshan Earthquake Archive” by Tangshan Archives. It was a record of this earthquake having the second largest number of deaths, including photos, files, video and audio recording material, etc.

After the 5•12 Wenchuan Earthquake, the national archives in the disaster-stricken area were busy on recovery work including collecting related materials, but not had any plan to construct similar digital archive like WEDA. Some of the related materials, such as web information about this major event were inevitably disappearing since created, which was obviously not included in the collection list of the national archives. Luckily, A LIS Professor from Public Administration School, Sichuan University, realized the importance of the digital preservation of this major event, and began to the creative thinking about the practical plan of WEDA construction.

In September 2008, four months after the 5•12 Wenchuan Earthquake, the Sichuan Federation of Social Science Circle (SFSSC) start an initiative to offer fund support to several research programs on earthquake recovery. The 2 LIS faculties of SCU and 4 librarians of SCU Libraries prepared a proposal to apply a program supporting the construction of WEDA and related research. This program was approved by SFSSC as one of the post-disaster reconstruction program, with 4,000 RMB fund support. Although without authority from national archives, library or any other government agency, WEDA construction went on going as a part of this research program at the beginning.

4.3 Solution of intellectual support

As we know, digital preservation involves multiple disciplines. With LIS faculties and librarians, the WEDA construction got professional guidance from library and information science, but still need other intellectual support from computer science, history and others related areas. After the communication with professionals in these areas, an expert of computer science and a postgraduate student of history were persuaded to join in this program. Then the team consisted of experts from LIS School, library, computer science and history, which gave enough intellectual support.

As a matter of fact, this multiple-discipline team performed well both in the program research and WEDA construction practice. The LIS faculties were familiar with digital preservation in theory and related research, so they tried to apply best theoretical output into the practice of WEDA construction, and always kept the practice to be coordinated with the original idea of the initiative. With rich experience in digital library, the computer expert helped a lot in the technical problems in digital preservation. The student from history contributed very nice ideas in the theoretical aspect of this program, promoting our understanding of social memory.

4.4 Solution of data resources

One of the most important things in the WEDA construction was the resource of the data. Because the materials the whole database based on had various types in content and form, the data resource was complicated and scattered from governmental agencies, non-governmental organizations and individuals.

Since SCU library was the largest university library in southwest China, and located in Sichuan, the province where the 5•12 Wenchuan Earthquake happened, it became an important data resource of WEDA, mainly in academic materials, such as academic books, periodicals, conference papers and dissertations. Furthermore, the local libraries and government departments in the earthquake stricken areas offered generous support in data resources. Meanwhile, we also seek cooperation with organizations such as the Red Cross Society of China, who gave us special authority to use the first-hand photographs in the WEDA construction after understanding the database will open to the public for free.

4.5 Solution of human resources

WEDA construction was started as a by-product of a research program. As we know, database construction is a hard task, which needs a great investment of human resources. In the view of the limited fund of this program, only 4,000 RMB, it became urgent to seek low-cost or even free human resources to do a lot of work in the process of WEDA construction. As the LIS faculties always have close connection with SCU libraries in different aspects and levels, and the earliest originator of this program, LIS Professor Yao, was also the chief director of SCU Libraries, 4 librarians from different departments of SCU libraries came to help after they understood the significance of WEDA construction. 16 LIS students, majored in Library Science, Information Science and Archival Science also volunteered. These volunteers not only completed task of data collection, arrangement and description, but also gave advice in how to construct this database.

4.6 Solution of standard and technology problem

In the process of WEDA construction, how to unify standards of metadata description and database management of different types of data became a problem. For digital preservation, there are some popular metadata standards such as DC metadata by Dublin Core Metadata Initiative, which was accepted by many institutions in the world. In China, there are some popular digital resources standards, such as Chinese Digital Library Standard (CDLS) by Institute of Scientific and Technical Information of China (ISTIC). In the academic library industry in China, there is a popular technical standard for digital library named “CADLIS.¹ Technical Standard” by the management center of China Academic Library & Information System (CALIS), which was widely used in many CALIS programs of academic libraries in China.

Given the condition of limited time and fund, the expert from computer science suggested that the main technical structure of WEDA can be built on the basis of CALIS Featured Database (CFD), like other featured databases Sichuan University Libraries had been constructed in the environment of digital library. As a result, WEDA followed the similar technical standard as CFD. Considering there was related

¹ CADLIS (China Academic Digital Library & Information System), a program funded by the National Project 211 in China, combined by two programs, CALIS (China Academic Library & Information System) and CADAL (China Academic Digital Associative Library).

archive, a different type of data from databases many libraries in China constructed seldom concluded before, related standards of archive and electronic records, such as Metadata for Electronic Records were followed too.

5. Achievement

After overcoming several obstacles and problems, this program was completed in June, 2009.

The output of this program is a 8,400-word research report, and the database “WEDA”, containing over 28,500 records, more than 80% of which are full-text data. The content of WEDA can be divided into five part as follows:

Name of Database	Type of Data	Format	Percent
5•12 Wenchuan Earthquake Record and Archive Database	Related records, archives and publications of the Central People’s Government of People’s Republic of China, and China’s local government including the government of earthquake-stricken area	Full-text	5%
5•12 Wenchuan Earthquake Academic Materials Database	Related books, newspaper reports, journal papers and dissertations in China and abroad	Full-text and Abstract	79%
5•12 Wenchuan Earthquake Web Information Database	Related webpages, blogs, bbs, etc.	Full-text	6%
5•12 Wenchuan Earthquake Picture Database	Related photographs	Full-text	7%
5•12 Wenchuan Earthquake Navigation Database	Navigations of related professional earthquake database and major events databases, professional earthquake organizations’ websites	Not full-text	3%

In 2010, the output of this program won the Social Sciences Achievements Award issued by the Sichuan Provincial Government.

6. Further development and future prospects

In 2011, based on the WEDA program, this team also applied a CALIS Featured Database Fund Program “Disaster Management and Crisis Response Database” (DMCRD). With more money support and rich experience, WEDA was expanded into DMCRD, a database containing both 5•12 Wenchuan Earthquake materials and other major event such as disaster and public crisis.

In 2012, two years after the Institution of Disaster Management and Reconstruction (IDMR) was set up at Sichuan University with the donation from Hong Kong, there is a plan to construct a Disaster Information Resource Center (DIRC), aiming at offer a Disaster Information Resource Database through a portal online to support the teaching and research of IDMR faculties and other scholars.

For the better development of WEDA and DMCRD, this team made a preliminary plan for the construction of DIRC for IDMR. In this plan, a large amount of WEDA data will open to the public for free to expand the database usage and its influence, especially in creating social memories of major events in digital age for next generation.

It is hoped that the further development of the WEDA program has an even brighter future.

7. Conclusion

In the practice of WEDA construction in China, several obstacles and problems from different levels and aspects has been overcome for creating social memories of major events. Though there is still a long way to go, the valuable experience we had in WEDA construction will be of a help to those in similar circumstances of struggling for the construction and preservation of digital documentary heritage.

ADDENDUM

Archiving large amounts of Individually-created Digital Content

Lessons from archiving the Occupy Movement¹

Howard Besser

Abstract

Archiving born-digital content from the “Occupy” movement can serve as a prototype for archiving all kinds of user-contributed content. In this paper, I discuss the tools and methods that a group of US organizations have identified for ingesting, preserving, and offering discovery services to large numbers of digital works where they haven't been able to really rely on the contributors to follow standards and metadata assignment. Topics covered range from automatic extraction of time-stamp and location metadata (and an empirical analysis of which upload services strip these out), to App development for uploading content along with permission forms, to maintaining lists of frequently-changing URL nodes for web-crawling, to issues in educating content creators in best practices. By getting involved in the creation stage of digital content, cultural organizations can better prepare for handling these works when they enter the repository at a later stage in their life-cycle.

Author

Howard Besser has been involved in building collections, services, and research directions for libraries and museums for almost 30 years, and began building image databases and linking cultural objects to maps 25 years ago. Currently, he directs New York University's Master's Degree program in Moving Image Archiving and Preservation. Previously, he was a Professor of Library & Information Studies at UCLA, where he taught and did research on multimedia, image databases, digital libraries, metadata standards, digital longevity, information literacy, distance learning, intellectual property, and the social and cultural impact of new information technologies. Besser has authored dozens of articles on automation and new research directions surrounding cultural materials.

1. Introduction

In the 20th century, memory institutions tended to focus on collecting content produced by organizations, and aesthetic objects created by skilled craftspeople (from poets and novelists to artists and designers). Records documenting the daily operations of an organization were considered important to save, whereas records documenting an individual's family (such as check-books, correspondence, family photographs, etc.) were only occasionally thought important enough to save. By the end of the 20th century there was a growing realization that relatively ephemeral works documenting everyday activity could, in aggregate, become critical works of enduring value, and would be valuable for reasons other than the original intention of their creator. When collected together, material such as amateur photographs and home movies offer important documentation of the daily lives of a particular ethnic group, or of rural life, or of

¹ An earlier version of this paper was presented as a Talk at the Coalition for Networked Information meeting in Washington on April 2, 2012, and will appear in a handout to accompany the 12th International Image and Research Conference in Girona, Catalunya, 20-23 November, 2012.

an urban ghetto at a particular point in time. And when these aggregations become part of a cultural institution's collections, they begin to transform scholarly research.²

With the ubiquity of digital recording devices (such as mobile phones) in the early 21st century, more and more of this sort of personal content is being produced by individuals and more and more of it is being shared with others. But it will not be possible for memory institutions to ingest the vast number of these born-digital works unless the process is at least partially automated. Cultural institutions will not have the resources to catalog and add metadata to each of the large number of works that will be coming from many thousands of sources (i.e., individuals).

As the InterPARES 2 Project demonstrated, to be successful, digital preservation must begin early in the lifecycle of a record/work; by the time it reaches the archive it is often too late. Using approaches developed during the Preserving Digital Public Television Project,³ New York City archivists and scholars have begun experimenting with collecting and preserving born-digital works captured or created by individuals associated with the "Occupy" movement in the US.

Archiving born-digital content from the "Occupy" movement can serve as a prototype for archiving all kinds of user-contributed content. In this paper, I discuss the tools and methods that a group of US organizations have identified for ingesting, preserving, and offering discovery services to large numbers of digital works where they haven't been able to really rely on the contributors to follow standards and metadata assignment. Topics covered will range from automatic extraction of time-stamp and location metadata (and an empirical analysis of which upload services strip these out), to App development for uploading content along with permission forms, to maintaining lists of frequently-changing URL nodes for web-crawling, to issues in educating content creators in best practices.

2. The Looming Problem with User-contributed information

Cultural heritage repositories such as libraries and archives have always collected material from organizations and individuals. But most of this material has entered the repository as part of a large collection that follows some kind of internal organizational and retrieval scheme (such as a newspaper's photo collection with its own specific numbering and metadata scheme, or an individual's file folders with labeling reflecting that individual's own personal logic). Libraries generally put this type of material in Special Collections departments to help indicate that each collection reflects the organizational schema of the donating individual or organization. And archives create a unique Finding Aid for each collection because each collection's organizational scheme is so different that the archive could not create an organizational scheme that worked across multiple collections.

In the late 20th century, as we moved more towards a world of born-digital collections, many challenges for cultural repositories (such as how to collect email correspondence, or how to organize the folders on someone's hard disk) emerged. But for the most part, a repository could expect that all the material coming from an organization or individual would follow a single organizational logic, and would

² Howard Besser, "Amateur Collections and Scholarship: Lessons from the impact of amateur collections on Cinema Studies research and on Film Archives' practice" (paper presented at Photo Archives and the Photographic Memory of Art History, Part III, Institute of Fine Arts, New York University, NY, NY, USA, 25-26 March 2011).

³ Howard Besser and Kara van Malssen, "Pushing Metadata Capture Upstream into the Content Production Process: Preliminary Studies of Public Television" (paper presented at DigCCurr 2007, Chapel Hill, NC, USA, 18-20 April 2007). <http://www.ils.unc.edu/digccurr2007/program.html>.

be encoded in a particular set of file formats (such as all documents in a particular version of Microsoft Word for each time period, and all digital movies in a particular version of Quicktime for each time period).

But many of the past artifacts collected by libraries and archives (such as letters and email and other types of correspondence) have been replaced in today's world by more atomized digital artifacts (such as blogs and social network postings) that are more connected to works originated by others. And in a world characterized by networked information and collaborative authorship, the cultural institution collecting material of historical and social importance will need to collect material coming from a variety of different sources, each with its own conception of metadata and file format standards.

Another important issue is the wealth of born-digital material that only exists on commercial services (such as gmail, Flickr, YouTube, or Vimeo). Many people believe that these services will preserve their material "forever". Few realize that many of these services quickly take something down with even the slightest challenge, and in no way should be considered long-term repositories. (See, for example, the website YouTomb, which tries to track takedown notices on YouTube.)⁴ And few realize that a number of the services assert ownership over content posted on them or require the signing of user agreements that prohibit many types of downloads or copying, making it technically illegal for a repository to copy material from a service, even with the original owner's permission. And few people realize that many of these services strip away important metadata in the file, making it difficult to sustain that file outside the environment of that particular service (see "Metadata loss During Upload or Download" study discussed below).

3. How Occupy material resembles what libraries and archives will face in the future

The material generated by the Occupy movement looks very much like the type of material that will be entering the archives and library special collections of the future. It is a vast quantity of user-generated everyday material, created by a multitude of different users.⁵ There is no easy way to control for quality, file format, or metadata. Unlike most organizational collections that try to enforce standards for metadata and file formats, there are not even guidelines suggesting what schemes should be followed. And because the content comes from so many individuals, it lacks even the semi-consistency that a single individual would apply to the items that he or she creates. And what might logically constitute a future "Occupy" media collection is actually found today spread over a multitude of commercial social networks (such as Flickr, YouTube, and Facebook) that each add their own organizational idiosyncrasies, and offer no guarantee that the material will remain posted for any length of time.

So, to preserve this type of material, we need to find smart ways to harvest metadata and analyze files, as well as to influence the behavior of potential contributors. A number of the methods that might be useful for future user-contributed collections were explored in the projects of the Activist Archivists, which are outlined later in this paper. Many of these methods are based upon the findings in prior projects on preserving born-digital material that Activist Archivists had participated in. From the InterPARES 2

⁴ As of this 2012 writing, MIT's Free Culture project has found that 9,760 videos were removed from YouTube for copyright violation, while 212,711 were taken down for "other reasons" (<http://youtomb.mit.edu/statistics>). Particularly for academic works that require a replicable citation, these private services cannot replace the functions of our traditional visual archives.

⁵ Besser, 2011.

project (2002-2006)⁶ we learned that if we hope to preserve electronic records, archivists need to be involved early in the life-cycle of that record, long before the record enters the archive. From the Preserving Digital Public Television project (2004-2010)⁷ we learned the effectiveness of automating metadata collection from the moment of first recording.⁸

4. The Occupy Movement and its Internet Presence

The Occupy movement began in September 2011 in New York City, and quickly spread to cities and towns across the US (and eventually to other parts of the world).⁹ The movement's key slogan—"We are the 99%"—reflects that the movement was fueled by a moral outrage at the control exerted on society by a small minority of the populace. The movement's name—"Occupy"—points to its tactic of "occupying" public physical spaces for 24 hours per day 7 days per week both to highlight the importance of those spaces to society's discourses, and to maintain a constant presence where people who pass by cannot help but notice the movement. This 24/7 presence in physical space also led to the development of self-organizing and community-building within the movement itself, and is reflected in the communal feeding of large numbers of participants (numbering in the hundreds, or in the case of NYC sometimes numbering in the thousands), and in collective providing of services for all participants (in the form of lending libraries, electrical power, wireless internet services, etc.).

In addition to physically occupying key public spaces, the movement engaged in extensive large-scale demonstrations involving thousands or tens of thousands of participants. Often these demonstrations highlighted what the movement saw as particular examples of systemic problems in society—the government's bail-out of financial firms (while not rescuing the worst-off individuals), the seizure of peoples' houses via foreclosure, etc. A major characteristic of the movement was the broad creativity shown in signs carried in protest marches, and in creative street-theater, where protestors would dress as bankers or governmental officials and act out satiric scenarios.

Like the Arab Spring movement that preceded it (and inspired it), the growth of the movement was fueled by communication mediated on the Internet. But, partially because of the high level of broadband Internet access and the ubiquity of smart-phones in the US, the number of digital photographs and video and audio recordings that movement participants posted online was astounding. Statistics from the photographic posting service Flickr show that 6 months after the first Occupy demonstration, more than half a million individual photos had been posted to this service with the tag of "#Occupy".¹⁰ Tens of thousands of individual videos were posted to YouTube during the first few months of the movement. By 6 months after the first Occupy action, 169,000 individual postings to YouTube had been tagged with

⁶ See http://www.interpares.org/ip2/ip2_index.cfm.

⁷ See <http://www.thirteen.org/ptvdigitalarchive/>.

⁸ Besser and van Malssen, 2007.

⁹ According to Wikipedia, by October 9, 2011 (3 weeks after its beginning in NYC), Occupy protests had taken place in over 95 cities across 82 countries, and over 600 communities in the US. (http://en.m.wikipedia.org/wiki/Occupy_movement, accessed August 19, 2012)

¹⁰ March 24, 2012 Flickr statistics show 632,089 items tagged with "#Occupy", 164,304 tagged with "Occupy Wall Street", 179,454 tagged with "Occupy Protest", 113,904 tagged with "#OWS", 40,572 tagged with "Occupy Movement", 27,202 tagged with "Occupy Oakland", and 9,164 tagged with "Zucotti Park" (location of the first NYC occupation).

“#Occupy”.¹¹ The vast amount of content created and the dissemination through commercial websites posed interesting problems for libraries and archives interested in preserving this material.

5. Activist Archivists

In response to the Occupy movement, in October 2011 students and recent graduates of NYU’s Moving Image Archive and Preservation Program (MIAP)¹² began efforts to explore the archiving and preservation of the media being generated by the Occupy movement. They felt that much of the spirit, decentralization, self-organization, playfulness, and whimsy of this protest movement would be lost to history if the media that documented this did not survive. Joined by MIAP Director Howard Besser, the group took on the name Activist Archivists, and began work on about a dozen different projects to archive the born-digital media content related to this movement,¹³ with most of the projects having potential impact on the archiving and preservation of all types of material that might be collected by cultural repositories in the future.

Many of the sub-projects involved collaboration with various partners. These included both collecting institutions (such as the NYU Library’s Tamiment Collection) and “working groups” from the Occupy Wall Street movement (including both the “Archives” working group, which mainly dealt with collecting non-digital artifacts such as posters and signs, and the “Media” working group).

It is important to note that certain predispositions of the Occupy movement may not be relevant to libraries and archives building collections from other sources. Those in the Occupy movement were very suspicious of conventional organizations, including universities and libraries, and often needed convincing that a conventional cultural institution might be a good repository for the artifacts that they created. Occupiers could also be characterized as having a “do-it-yourself” (DIY) mentality, not wanting to rely on professionals outside their community to organize and provide access to the material. This was part of a critique of conventional media dissemination outlets which did not do a good job of explaining the movement, and appeared to manipulate news coverage. The Occupiers wanted to control their own story. Their ideology also made them suspicious of any type of exclusive arrangement, including giving their material only to a particular repository. And their consensus decision-making process made it difficult for a repository to try to come to an agreement with the group, as a group discussion on a topic such as this might range over several meetings, and each meeting might be composed of a slightly different group of participants (and discussion from previous meetings had to be repeated to and accepted by the newcomers).

6. Activist Archivists Projects

Cultural institutions speak with pride about their “special collections”—where they are the only location housing particular “original” objects or records. This kind of exclusivity runs counter to principles of

¹¹ March 24, 2012 YouTube statistics show 169,000 items tagged with “#Occupy”, 98,400 tagged with “Occupy Wall Street”, 70,500 tagged with “Occupy Protest”, 50,300 tagged with “#OWS”, 54,800 tagged with “Occupy Movement”, 13,400 tagged with “Occupy Oakland”, and 6,690 tagged with “Zucotti Park” (location of the first NYC occupation).

¹² See <http://www.nyu.edu/tisch/preservation/>.

¹³ See <http://www.activist-archivists.org/>.

openness and sharing which are dear to the Occupy movement. This, coupled with the fact that many collecting institutions are part of larger organizations that Occupiers associate with Wall Street and real estate development (such as NYU) made Occupiers hesitant to sign the standard donor agreements that collecting institutions normally require. When New York collecting institutions were having trouble convincing Occupiers to execute donor agreements and hand over their material, Activist Archivists began promoting **Creative Commons licenses**. If these licenses were properly executed by content recorders/creators, they would allow any repository to copy, collect, and disseminate the material for the purposes of research, education, and preservation. As opposed to the exclusivity of standard donor agreements, Creative Commons licenses were much more in line with Occupy principles.

Because many in the Occupy movement did not immediately recognize the value of saving artifacts representing their movement, one of the first projects undertaken by the Activist Archivists was to create a short simple postcard that succinctly explained why saving these artifacts should be important to the movement. Much effort was made to explain this using the value system of the movement, and significant effort went in to avoiding what many in the movement referred to as “outsider language”. The “**Why Archive**” postcard briefly explained that saving this material would serve values dear to the movement, including: Accountability, Self-Determination, Sharing, Education, and Providing Continuity. It explained that it was important to “Record and Collection” what was happening in the movement in order to “Preserve the record”.

Another important Activist Archivist project sought to provide advice to those recording Occupy events. The “**7 Tips to Ensure our Video is Usable in the Long Term**” offers important advice to those recording video, still images or sound. Advice includes: collecting details about the recording, keeping the original raw unaltered footage, making the recording discoverable/accessible, contextualizing it, making it verifiable, allowing others to collect and archive the recording, or being very careful about archiving it yourself. (For this last point, the Tips suggests looking at the Library of Congress’ website on Personal Archiving.) Ideas from these “7 Tips” were expanded into a much lengthier set of advice in “**Best Practices for Video Activists**.” This document also discussed legal restrictions involving getting permission from those you record (which in the US differs depending on which state the recording takes place in), as well as issues involving copyright (and stresses the idea of executing a Creative Commons license which will allow a repository to archive the material and make it available in the future).

An Activist Archivist research project revealed significant issues for any organization seeking to preserve digital videos posted to commercial websites. MIAP student Rufus de Rham’s “**Metadata Loss During Upload or Download**” empirical study uploaded digital video files from 3 different consumer recording devices (iPhone, Android, and Canon t2i) to 4 different internet dissemination services (YouTube, the Internet Archive, and both Vimeo’s paid and free services). He then examined over 200 fields of metadata present in the original files, and found that YouTube and Vimeo’s free service stripped out almost all the metadata, and only the Internet Archive’s service maintained all the metadata intact. Important metadata such as date, time, and GPS location were stripped out of the file header by YouTube and Vimeo’s free service, and the files posted on these two services were heavily compressed. This means that cultural organizations seeking to collect videos posted to online services would have to reconstruct date, time, and location information for anything they collected from YouTube or Vimeo’s free service, and would also be collecting poor quality videos without assurances of fixity, integrity, or authenticity.

Activist Archivists also worked with the Internet Archive’s Archive-It team to help identify the frequently-changing URLs that would serve as nodes for **crawling and preserving Occupy websites**.

Since the beginning of their collaboration with Activist Archivists, Archive-It has maintained a history of the changing scope of Occupy websites.¹⁴

Activist Archivists also worked on a “**Best Practices for Content Collectors.**” This document aimed to pull together information gleaned from other Activist Archivist projects, and to offer advice to cultural repositories on how best to collect this type of content. It suggested that with content pulled from the Internet Archive, metadata for date and time recorded (and possibly geographic location as well) could be automatically extracted from metadata in the file headers. It discussed some Occupiers’ sensitivity around traditional organizations and exclusivity, and suggested ways in which an archive could collect materials without being tied to an exclusive agreement. It discussed the sensitivity of some protesters around security, and offered suggestions for using software tools that could automatically detect faces and place small black bands over peoples’ eyes in order to make individual identification difficult. When drafts of this document were distributed to managers of existing collections asking them to review and comment, none responded. So this document was never finalized and released.

Two Activist Archivist members worked closely with NYU’s Tamiment archival collection, which had arranged for the **digital audio recording of daily 2-hour “Think Tank” discussions of strategies and tactics** by Occupy Wall Street participants. The Activist Archivists suggested ways in which important metadata could be automatically extracted from the recordings, as well as working on the spreadsheets of metadata related to each recording. Guidelines for the recordings insisted that key metadata be captured in redundant ways. So, for example, the date that should have been captured by the Zoom H2n digital audio recording device was also embedded within the file-name, and was supposed to be orally recited as part of an initial script that was read into the recorder at the start of each session. This redundancy was important because the recording device’s date and time were not necessarily reset when the device’s battery ran out of energy, and the oral script was not always read at the beginning of the recording. This form of redundancy would also protect against a future time when the metadata might become separated from the content. And the reading of the same exact script at the beginning of each session would allow for future, more sophisticated speech recognition software to automatically extract important metadata. The script also is designed to systematically record information including who is responsible for the recording, what is expected to be discussed in this session, and that participants desire to execute a Creative Commons license that will allow archives to preserve and disseminate the recording.

Activist Archivists also advised NYU’s Tamiment archival collection on **preserving digital videos posted on YouTube**. Selection of which videos to capture and preserve had been a tedious process, with the Tamiment Director viewing each YouTube video, and deciding whether or not it was worthy of preservation. This process would not scale to the 100,00 videos that had been posted to YouTube tagged with “Occupy Wall Street” just 6 months into the movement. Activist Archivists suggested that categories for the videos be developed (celebrity visits; internal workings such as the library, kitchen, and media; confrontations with police; labor union involvement; housing/foreclosure protests; etc.), and that selection of the most important videos be crowd-sourced to members of the Occupy Wall Street Working Groups (both Archives and Media). Each participant could fill out a web form indicating what they thought were the 5 most important YouTube videos to preserve in each category. This method would scale up as the number of videos posted to YouTube increased, and it was more in line with the participatory nature of the Occupy movement. This method has still not been implemented due to the illness and eventual death of the Tamiment Director.

¹⁴ See <http://www.archive-it.org/collections/2950>.

As the Occupy movement approached its one-year anniversary, Activist Archivists turned their attention to other collections and movements. At the time of this writing (August 2012) they are planning work sessions with small independent political archives, as well as following up on problems that arose with Occupy material that proved technically very challenging (such as the capture of streaming media from services such as LiveStream).

Another technically-challenging project that the Activist Archivists never pursued involved creating Apps for various recording devices that would allow someone recording events to pre-populate metadata to accompany their recordings. Such an App could allow the recordist to execute a Creative Commons license, embed their own name or a pseudonym as metadata into the file header, and add date/time and GPS location metadata to the digital object. It would also allow the recordist to choose which (multiple) sites on which they want to post their recording. Using this App, a person could just record something, pull up the App, either check “OK” or change some parameters, then push a button, and the result would be that the recording with extensive metadata would be sent to all relevant online services. This would also solve the problem that many recordists want to post a video on YouTube because of its wide dissemination, but a copy could also go to the Internet Archive where the metadata would not be stripped out and where it would much more likely to survive over time. Some pieces of this idea have been incorporated into the InformaCam plugin for ObsuraCam as a collaboration between the human rights group Witness and the Guardian Project.¹⁵

7. Conclusion

Collecting and preserving born-digital content from a wide variety of individual sources is a looming problem that collecting institutions will need to face in the near future. Archiving born-digital content from the “Occupy” movement can serve as a prototype for archiving all kinds of user-contributed content. As this paper has shown, Activist Archivists tackled a wide variety of problems inherent in selecting, capturing, and preserving media related to the Occupy movement. They explored various methods for convincing the individuals who record these events to make sure that their recording devices were set to automatically record date, time and location; to use consistent metadata and file-formats; to post to sites that will not throw away their metadata; and to execute creative commons licenses that will give cultural repositories the legal right to preserve and make available these recordings. They also advised collecting institutions on how to articulate their need to preserve to the individuals that make recordings, and provided advice on both redundant methods of capturing metadata for recordings, and on scaling the selection process in creative ways. These efforts are likely to prove useful for solving a problem that most collections will face in the near future—how to organize, preserve, and provide access to that large amount of user-generated content that most collections will receive in the future (and not have the time to catalog or convert).

¹⁵ See <http://blog.witness.org/2012/02/introducing-informacam-the-next-release-of-the-securesmartcam-project/>.

Digitisation of Small Sound Collections

Problems and Solutions

Nadja Wallaszkovits

Phonogrammarchiv, Austrian Academy of Sciences

Abstract

The paper discusses digitisation of small sound collections, focusing on the structural, technical and conceptual problems of practical implementation and realization, such as: assessment of the collection and its state of preservation; assessment of required and available equipment; development of a preservation plan; proposal of a prioritized sequence of actions based on different urgencies for different parts of the collection; definition of equipment needed (test, analogue, digital) and design of a business plan of investment; training of the local staff in digital audio archiving techniques and methodology; installation of equipment and initiation of work; subsequent technical and conceptual support. The discussion also includes the implementation of an open source based database and server system, which can be individually adapted and expanded.

Author

Nadja Wallaszkovits studied musicology and audio engineering in Vienna and has been working as sound engineer for international recording companies. In 1998, she joined the Phonogrammarchiv, where she manages the audio department as a specialist for audio restoration, rerecording and digital archiving. She is consultant for archival technology and has held several international training seminars. She works as guest lecturer at the University of Vienna and at the University of Applied Sciences HTW in Berlin. Nadja Wallaszkovits is vice chair of the IASA Technical Committee and vice chair Central Europe of the Audio Engineering Society (AES).

1. Introduction

Over the past decades, research institutes all over the world have accumulated significant collections of linguistic, ethnomusicological and folkloristic audio materials. In many cases these holdings represent unique sound and audiovisual documents of remarkable cultural value and will only survive if transferred into the digital domain in the mid-term. The Phonogrammarchiv has been involved in several such digitisation projects of small scale holdings, which have partly been funded from outside.

One of the starting points for the Phonogrammarchiv's engagement was the political change around 1990 in former European Eastern Bloc states. Since the 1950s, research traditions in these countries had produced, as compared to research institutions in the West, an unusual amount of mainly audio documents in the fields of linguistics, ethnomusicology and cultural anthropology at large. The political changes were accompanied by severe changes in the administration and funding of heritage and research institutions, which partly put these collections at serious risk. To explore the situation, audiovisual archives have been systematically contacted and visited by representatives of the Phonogrammarchiv between 1992 and 1995.¹ An actual overview of the situation of audiovisual research collections and their preservation status can be found in the report published in the framework of TAPE (Training for

¹ This survey has been funded by the Austrian Ministry of Science and Research.

Audiovisual Preservation in Europe).² These contacts initiated a number of smaller and more substantial contacts, of which several ultimately lead to digitisation projects.

The projects supported by the Phonogrammarchiv concerning strategic planning and practical implementation of digitisation include the collections of the Institutul de Etnografie si Folklor “Constantin Brăiloiu” of the Rumanian Academy of Sciences, Bucharest (Rumania),³ the Institute for Cultural Anthropology and Art Studies (formerly Institute for Folk Culture) of the Albanian Academy of Sciences, Tirana (Albania)⁴ and the Phonogrammarchiv St. Petersburg, Pushkinsky Dom, (Russia).⁵ In 2007, the Phonogrammarchiv of the Austrian Academy of Sciences was the UNESCO Jikji Prize winner. In the application for the prize, it had been stated that the prize money (donated by Cheongju City) would be used to contribute to safeguarding an audiovisual collection, preferably from a country with developing economy. In the same year the José Maceda Collection of the University of the Philippines was inscribed by UNESCO on the International Register of the Memory of the World Programme. Familiar with the collection and its precarious situation from previous contacts, the Phonogrammarchiv thus decided to spend the prize money on the digitisation of this important corpus of ethnomusicological recordings.⁶ Recent supporting projects of the Phonogrammarchiv include the digitisation of the sound archive collection of the National Archives and Library of Ethiopia, partly financed by UNESCO.

2. Similarities and Common Problems

In many cases the collections have been created by collectors or researchers privately or in cooperation with an institute, and have been held and stored outside audiovisual archives proper. At some point such documents were incorporated to the archival holdings of research institutes or other somehow related archives. Latest until the material should be accessed, it creates a bundle of problems:

Usually the keeping institution is not an A/V archive in a narrower sense and therefore suffers from the lack of modern professional replay facilities and A/V media knowledge. Additionally, staff shortage, inadequate funds and sometimes management problems (e.g., awareness about the vulnerability of A/V media, etc.), as well as missing expertise and lacking financial means to keep the analogue originals and subsequently the digital data alive, sum up.

Typically such collections are “mixed media” holdings. The big diversity of archival holdings is on one hand caused by such incorporations of external collections as described above, but sometimes also by a changing or missing collection policy.

In the majority of cases the holdings have been created with (many times privately owned) consumer equipment. The collections have been recorded and stored under irregular and/or irreproducible conditions. This generates a wide range of associated challenges concerning transfer and digitisation planning: Usually, the lack of (technical, but also other, e.g., content related) documentation (e.g., information about recording formats and parameters, like speed, trackwidth, etc.) makes it difficult to

² Dietrich Schüller, “Audiovisual research collections and their preservation,” European Commission on Preservation and Access, 2008. Report published in the framework of TAPE (Training for Audiovisual Preservation in Europe). http://www.tape-online.net/docs/audiovisual_research_collections.pdf.

³ Financed within the EU-funded project “ethnoArc” (2006-2008).

⁴ Financed by the Austrian Development Agency (ADA) (2005-2009).

⁵ Financed by two INTAS Projects and most recently by the Endangered Archives Program of the British Library.

⁶ For further information, see http://www.phonogrammarchiv.at/wwwnew/jose_maceda_coll_e.htm and <http://portal.unesco.org/ci/en/files/30579/12767596313Report-useprize-2007.pdf/Report-useprize-2007.pdf>.

accurately estimate the quantity of the holdings and time and expenses needed for digitisation. This easily results in miscalculation of digitisation efforts and costs. Format obsolescence is an additional problem, especially if proprietary (consumer) technology has been used for recording. The preservation status of the holdings can be critical and therefore the time factor for transferring the collection can vary strongly.

Summarising all these factors, such collections are, in many cases, of unique contents, containing mostly field recordings, which are incorporated in a research environment. The heterogeneous structure and the possibly critical preservation status usually require an individual manual transfer and adequate scientific documentation, but infrastructure, expertise and financial means are lacking.

The challenge is to implement a manual small scale and high quality transfer, meeting the actual technical standards: IASA TC03⁷ and TC04,⁸ at low costs and with high reliability, incorporating infrastructure and technical skills.

3. Preconditions

Such projects need a number of basic decisions to be made, so a detailed knowledge of the general situation of the institute and the collections to be digitised is necessary. The better the collections are documented, the easier their proper digitisation can be planned and cost effectively implemented.

One basis decision is the question of outsourcing the digitisation versus doing the job in-house. More and more archives are outsourcing (parts of) their preservation/digitisation work. This major decision demands appropriate choice of a dedicated service provider, as well as ongoing quality control, to guarantee accurate transfers. One of the burning problems for an archivist proper is: How can we ensure that the work paid for is done to archival perfection? Especially transferring large audio archives represents a daunting task, but also small scale migration projects have to verify the quality of their outsourced digitisation projects. Digitisation, if done properly, is a quite complex process and faces lots of possible error sources, which are not easy to trace. Additionally, there are lots of commercial (small scale) providers, and very often, there has to be made a compromise between costs and quality. Usually, such compromises are decided in favor of costs. In doing so, audio digitisation quality usually becomes a minor factor—therefore the quality of the digital original file and all future copies will be of bad quality. The questions of how to find a trustworthy provider and of how to control its work represent a special challenge and need very high skills.

In many cases, it is useful to think about an in-house transfer. Basis requirements that make an individual in-house transfer effective are:

A minimum size of the collection

If the collection is too small, cost efficiency is not given. In this case it is recommended to seek cooperation with institutions in similar situations. In practice it is problematic to accomplish such a joint venture due to lack of a strong interregional management, mistrust and internal competitiveness. As an alternative outsourcing could be a solution. Unfortunately it is problematic to find an adequate and trustworthy service provider at moderate costs, especially if the collection should not be shipped across the globe.

⁷ Schüller, “Audiovisual research collections and their preservation.”

⁸ IASA Technical Committee: Standards, Recommended Practices and Strategies, “IASA-TC 03: The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy,” version 3, December 2005. Published in English, German, French, Swedish, Spanish, Italian, Russian and Chinese. http://www.iasa-web.org/downloads/publications/TC03_English.pdf.

From our experience, a point of intersection is given at a minimum number of about 2000 hours of audio material to be digitised.

Expected increase of the collection

If the size of the collection is very small, but steady increase is to be expected, the setup of an in-house infrastructure is reasonable as well. The implementation of a center of A/V competence can be a big advantage for other institutions and also a financial benefit for the institute running the facilities.

Not too many different (multimedia) formats to cover

In case of a large variety of different formats (e.g., very small quantities of a lot of different and/or obsolete media), setting up an in-house infrastructure will also be very expensive and therefore not very effective. In this case, it should be considered to outsource, e.g., parts of the collection for digitisation, keeping in mind the associated problems, as mentioned above. In practice, a mixed strategy could be taken into consideration, if necessary.

In combination with the fact that educated (scientific and/or technical) staff with knowledge of the collection should be available, and cooperation with local IT specialists is possible, these requirements make an individual, small scale, in-house transfer cost-effective and successfully implementable.

4. Assessment and Appraisal of the Collection and Infrastructure

The first step in starting such a digitisation project is an assessment of the collection with the aim to gain as much information as possible about the overall preservation status, the size in terms of playing time and technical parameters required for calculation of replay equipment. It is a matter of fact that the more information about the collection is available, the better the calculation of needs can be carried out.

One possibility to gather information about the situation of the individual collection is to send out a questionnaire covering the important main topics like amount of holdings listed by type of carriers, information about storage area, analogue and digital infrastructure, IT implementations, metadata infrastructure, and so on.

A12								
LIBRARY #	RB	MATERIAL	THICKNESS MIL	SIDE	HEAD IN (HI) / TAIL OUT (TO)	SPEED	GOOD / BAD	CONDITION
IL1 69 / SCB1	5	POLYESTER	1.5	A&B		7 1/2		START OF MUSIC WAS CUT
IL2 70	3	ACETATE	1.0	(L) MONO		3 1/4		CURLED
IL3 70	3	ACETATE	1.0	(L) MONO		3 1/4	✓	CURLED
IL4 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL5 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL6 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL7 71	5	POLYESTER	1.6	(L) MONO		3 1/4	✓	
IL8 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL9 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL10 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL11 71	5	POLYESTER	1.5	A&B		1 1/2	✓	
IL12 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL13 71	5	POLYESTER	1.5	A&B		1 1/2	✓	
IL14 71	5	POLYESTER	1.5	A&B		1 1/2	✓	
IL15 71	5	POLYESTER	1.5	A&B		1 1/2	✓	
IL16 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL17 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL18 71	5	POLYESTER	1.5	A&B		A 1 1/2 B 3 1/4	✓	
IL19 71	5	POLYESTER	1.5	A&B		A 1 1/2 B 3 1/4	✓	
IL20 71	5	POLYESTER	1.5	A&B		1 1/2	✓	
IL21 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL22 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL23 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL24 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL25 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL26 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL27 71	5	POLYESTER	1.5	A&B		1 1/2	✓	
IL28 71	5	POLYESTER	1.5	MONO		A 1 1/2 B 3 1/4	✓	
IL29 71	5	POLYESTER	1.5	A&B		A 1 1/2 B 3 1/4	✓	
IL30 71	5	POLYESTER	1.5	MONO		1 1/2	✓	
IL31 71	5	POLYESTER	1.5	A&B		3 1/4	✓	
IL32 71	5	POLYESTER	1.5	A&B		3 1/4	✓	

Figure 1. Result of the questionnaire, José Maceda Collection of the University of the Philippines. Illustration: N. Wallaszkovits.

IASA offers a special publication to examine the issues underlying the process of setting priorities for the digital transfer of analogue and digital audio content, and to deliver a statement of principles for use by sound archives in their planning for digitisation. This is the “Task Force to establish Selection Criteria of Analogue and Digital Audio Contents for Transfer to Data Formats for Preservation Purposes”⁹

A very helpful tool, especially for estimation of the overall preservation status of individual collections, has been developed by the Indiana University Digital Library Program within the Project Sound Directions. The *FACET* (Field Audio Collection Evaluation Tool) can be downloaded from their website.¹⁰ This institution furthermore provides a very useful reader for best practices on audiovisual archiving, focusing on the individual transfer of heritage material.¹¹

Other possibly helpful tools for such an assessment have been developed within the project PrestoSpace, the preservation, storage and cost calculators.¹² Although basically designed to meet large scale approaches, these tools can be used in small scale approaches, too.

To gain a more detailed impression of the general status of the collection, it is advisable to take random samples from the collection and have a more detailed analysis of the preservation status and recording parameters. Ideally, adequate replay-facilities are available to carry out such detailed checks.

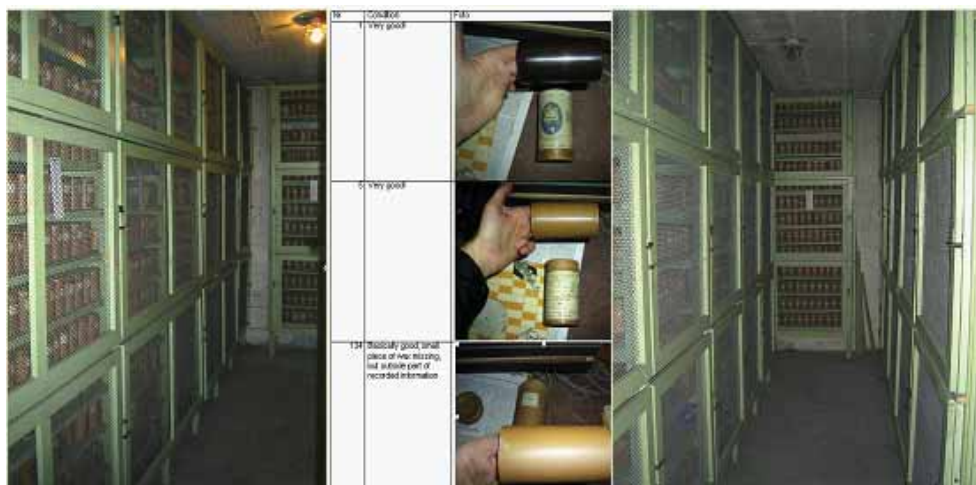


Figure 2. Assessment of the cylinder collection of the Institutul de Etnografie si Folklor “Constantin Brăiloiu” of the Rumanian Academy of Sciences, Bucharest (Rumania). Photograph: N. Wallaszkovits.

Additionally, it is useful to make an assessment of the equipment that has been used for recording of the original tapes. Although in many cases it is not possible to get information about all machines, it can be still helpful to find out details about track formats, speeds and other—maybe irregular—parameters to expect within a collection. This helps to avoid miscalculation concerning playing time and equipment

⁹ International Association of Sound and Audiovisual Archives, “Task Force to establish Selection Criteria of Analogue and Digital Audio Contents for Transfer to Data Formats for Preservation Purposes,” 2004. <http://www.iasa-web.org/downloads/publications/taskforce.pdf>.

¹⁰ See <http://www.dlib.indiana.edu/projects/sounddirections/facet/downloads.shtml>.

¹¹ Mike Casey and Bruce Gordon, eds, *Sound Directions: Best Practices for Audio Preservation*, Indiana University, 2007. http://www.dlib.indiana.edu/projects/sounddirections/papersPresent/sd_bp_07.pdf.

¹² See <http://www.prestocentre.org/library/archival-storage/tools>.

needs. In a further step, the available equipment (including the IT infrastructure) should be evaluated for usefulness in the digitisation and documentation process.

The storage area and building constructions should also be assessed, to calculate the measures needed for optimising long term storage. Sometimes already minor equipment acquisition, like the purchase of a hygrometer and a dehumidifier can significantly improve long term storage conditions.

An assessment of the existing metadata structure is also useful and will help calculating costs for database implementation.



Figure 3. Assessment of the metadata structure and of the original recording equipment of the collection of the Institutul de Etnografie si Folklor “Constantin Brăiloiu,” Rumanian Academy of Sciences, Bucharest (Rumania).
Photograph: N. Wallaszkovits.

5. Developing a Preservation Plan

In the next step a preservation plan has to be developed, proposing a prioritised sequence of actions, based on different urgencies for different parts of the collection. In such a preservation plan the focus is set on the most endangered medium with the highest scientific value and the most frequent access, balancing these factors carefully. The preservation plan should include calculations for optimising storage conditions, transferring to digital, the definition of equipment needed and finally should represent a profound basis for designing a business plan of investment.

A setup of infrastructure should include calculations and specifications for:

- Analogue replay equipment
- A/D Converter
- Maintenance equipment and spare parts
- Digitisation workstations
- Access stations
- Server
- Database
- Additional needs and infrastructure
- Running costs to keep digital data alive

The preservation plan developed in our projects is based on a concept for a small scale approach to digital storage and follows the specifications and recommendations of IASA-TC 03¹³ and IASA TC04.¹⁴

The simplest and smallest concept is a single operator input and access station, where the archival files and the access copies are created. The digital audio workstation (DAW) has a desktop RAID array attached. The contents of the RAID have to be regularly and at least backed up to data tapes (e.g., LTO). The data tapes are manually loaded for storage on traditional shelving, with increased attention to minimising the presence of dust and other particulates and pollutants. As disc storage has become constantly cheaper during the last years, parallel copies on (externally stored) hard discs will optimise the data security. This concept is a comparatively easy, reliable and applicable solution. This setup requires a well structured plan for digitisation, as well as careful management of copy location information and version information, which is done semi-manually.

The access to the material can be managed traditionally, via copies on access media (like CDR or DVDR), or by browsing copies (e.g., in MP3 format) that are accessible from the workstation via a user account. Ideally, access is separated from the ingest station, so this approach is only applicable in very small institutions, following a restrictive user policy. On-line access is not included, as this requires additional functionality.

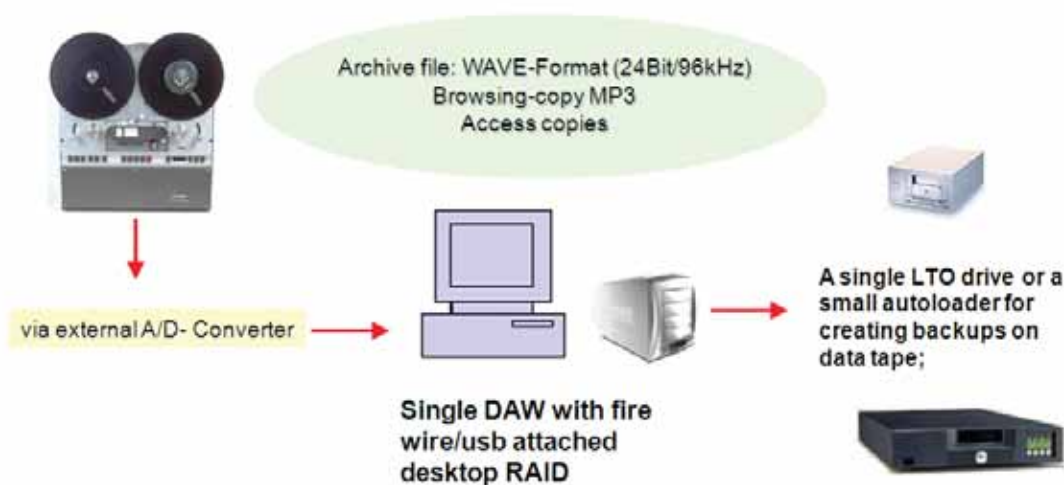


Figure 4. Simple concept for a small scale approach to audio digitization. Illustration: N. Wallaszkovits.

The system can be expanded to a small scale network for two or more ingest stations, also of different media (e.g., cassettes, DAT, etc.) and one or more users, on basis of a Network Attached Storage (NAS) system with a capacity between ~ 0.5 to 20 terabyte (TB) of disk storage. In combination with an LTO autoloader this is already a midrange solution but certainly needs a higher level of administration to work properly.

Whenever such manually handled solutions come into consideration (in most cases due to lack of money), a stringent copy and safety strategy has to be implemented. This can only be reached by using

¹³ IASA Technical Committee: Standards, Recommended Practices and Strategies, "IASA-TC 03."

¹⁴ IASA Technical Committee: Standards, Recommended Practices and Strategies, "IASA-TC 04 Guidelines on the Production and Preservation of Digital Audio Objects," 2nd ed., 2009. http://www.iasa-web.org/special_publications.asp.

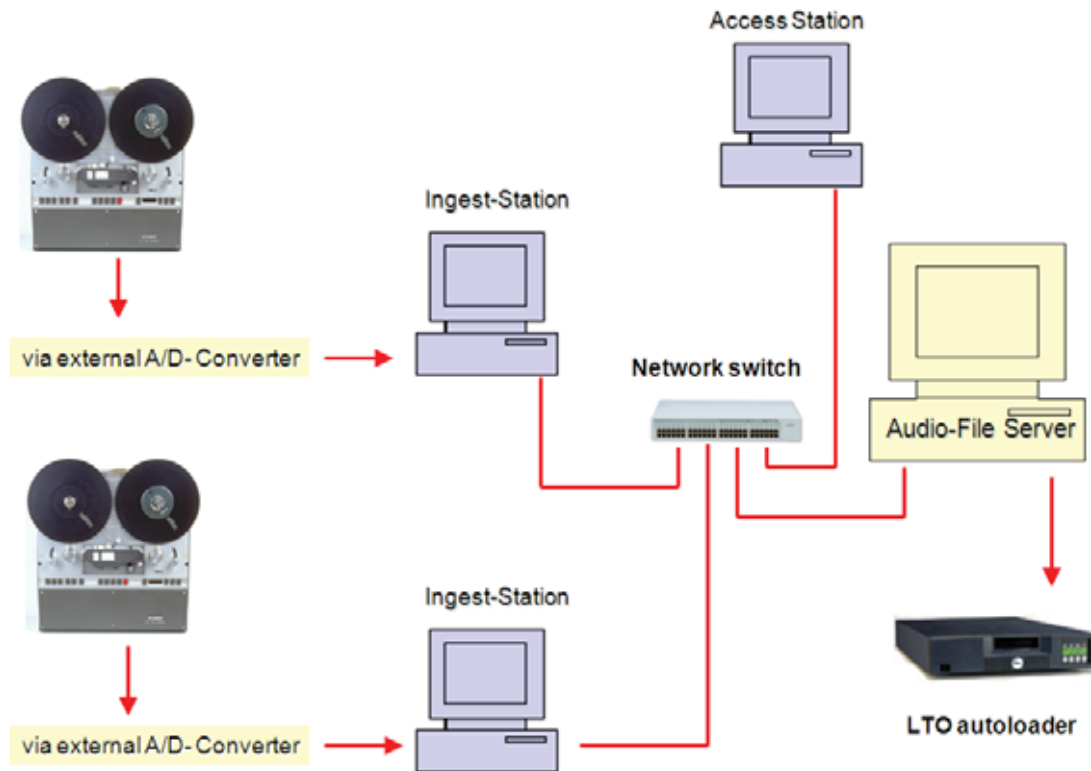


Figure 5. Small scale network for two ingest stations and one or more users. Illustration: N. Wallaszkovits.

unique identifiers and that can be written to the data tapes header and can be useful for data verification, especially in future migration processes.

It is beneficial to direct the acquisition of all kind of necessary equipment in a way that long term support and service can be managed locally. Although it can be time consuming and long-winded to find a local dealer for dedicated equipment, this concept allows efficient troubleshooting and time saving in the long-term.

To ensure the maintenance of such a concept, support is managed by cooperation with local IT specialists (e.g., coming from the local Academies IT departments). It is necessary to have the assurance that the system will be regularly updated and that the upkeeping of the digital data is guaranteed. A digital long term storage strategy following the Open Archival Information System (OAIS)¹⁵ should be followed preferably.

6. Metadata

As dedicated tools for capturing metadata were missing in the infrastructure of our digitisation projects, we had to find a cost efficient and easily implementable solution for this important point. The focus was set on a concept that can be easily locally managed and hosted, can be easily integrated to an existing

¹⁵ CCSDS 650.0-B-1, "Reference Model for an Open Archival Information System (OAIS)," Blue Book. Issue 1. January 2002, adopted as ISO 14721:2003. http://nssdc.gsfc.nasa.gov/nost/isoas/ref_model.html and <http://public.ccsds.org/publications/archive/650x0b1.pdf>.

intranet, can be opened to the Internet if necessary, and provides very good safety mechanisms and data security strategies. In cooperation with our consultant, Dr. Andreas Matzke, who is working for the Senior Expert Service SES in Bonn, Germany, a database which is based on widely used open source software and kept very simple and easygoing, so that it can be handled by untrained staff, was developed. The system relies on widely used open source software (Linux Ubuntu, database system MySQL, Apache Server, user interface based on php5).



Figures 6a) and b). The database of the Institute for Cultural Anthropology and Art Studies of the Albanian Academy of Sciences in Tirana (Albania). **a)** (left): Screenshot of the search page of the database **b)** (right): Screenshot of the database input module. Illustrations & design: Andreas Matzke.

The database connects documents and according metadata, supports the collection of documents and metadata and offers facilities to access its content. One of the basic design concepts of the system was, to make handling and maintenance as safe and easy as possible. In fact, the number of screen pages used as an interface between the archivist and the system is rather small, and there are just only three different types of screen pages a normal user of the archive will see. As a result using the system can be learned in a short time. This should not keep the archivist from taking special care during the manipulation of metadata. The archive system can easily be adapted to future needs, like additional types of metadata, different and complex retrieval functions, integration of external existing documents and data or facilities to transfer its own content to other external archives. A specific feature is also its capability to handle multiple scripts. It has been already used in Arabic applications, and in the Ethiopian Project it will use Amharic script.

7. Training And Practical Implementation

Within the projects, a two weeks hands-on training of technically interested/ educated staff (with practical archival experience) in the Phonogrammarchiv Vienna was calculated. The main focus of these trainings was set on the unmodified high quality transfer (conform with IASA TC03/TC04¹⁶), digitisation workflow, maintenance of analogue tape machines and documentation, especially of transfer metadata. Our experience shows, that this is a good time span for training, balancing costs and demands, provided that the staff to be trained has a minimum preparatory education.

¹⁶ IASA Technical Committee: Standards, Recommended Practices and Strategies. "IASA-TC 03" and "IASA TC04."

The practical implementation of the projects includes the acquisition of adequate replay equipment. In practice, as new analogue magnetic tape machines meeting the IASA specifications are not available from the market anymore, used replay equipment has been purchased and revised to fit the necessary specifications and parameters outlined in IASA TC04.¹⁷ The equipment was shipped and thereafter was installed locally. Digital equipment and other infrastructural devices have been purchased from local providers, mainly to ensure local maintenance and support. After successful on-site installation, the local staff was trained in house and a digitisation of some first analogue tapes was carried out. This task was calculated with up to three weeks for server and database installation, and one week for the audio setup. Within this process the workflow for digitisation has been optimized and adapted to the needs of the archives specifically.

After successful processing of some critical tapes, the first results have been presented to the department heads. Additionally, the projects have been regularly presented to the public by organising press conferences and digitisation workshops. By means of such specific actions, public awareness could be raised and the importance of the collections as a promising basis for a sustainable consolidation of such repositories of important ethnomusicological sources could be emphasised.

8. Subsequent Technical and Conceptual Support

The projects are meanwhile autonomously working or already finished (except the digitisation of the sound archive collection of the National Archives and Library of Ethiopia, which is still in the phase of preservation planning and equipment acquisition).

Nevertheless, the Phonogrammarchiv gives subsequently support and training concerning A/D transfer, technical problems on the playback side, as well as help in obtaining spare parts and additional equipment to guarantee an individual high quality transfer with optimum signal retrieval from original tapes. Additional features of specific quality checks have been implemented, so that the quality can be controlled from afar. Long-term service of storage facilities is solved by cooperation with local IT specialists, as outlined before. Additionally, a second digital copy of the important collection of the Institute for Folk Culture of the Albanian Academy of Sciences has been deposited in parallel in the Vienna Phonogrammarchiv. This acts mainly as an additional safety measure for ultimate disaster preparedness, but also as a second access point under restricted access rights.

Conclusion

The examples of practical implementation of small scale digitisation projects outlined above show, that the IASA guidelines and standards related to the long term preservation and digitisation of sound recordings are a most useful reference in daily archival work. Nevertheless, there are always enough lessons to learn in practical application. Creativity and helping yourself in situations of lacking infrastructure, a soldering iron and the famous gaffer tape, as well as good mood and the enthusiasm to have the job done, are essential ingredients and make life easier. Balancing between different mentalities is important, and sometimes the ambition to meet the ultimate quality requirements has to be

¹⁷ IASA Technical Committee: Standards, Recommended Practices and Strategies. "IASA-TC 04," 4, p. 8f. and p. 32ff.

pragmatically modified and efficiently realised. The outstanding personal engagement and enthusiasm of the staff is the basis for successful implementation.

The digitisation projects discussed in this paper have been also designed as regional pilot projects that should be seen in a larger context. They present a strong relation to identity, democracy and human rights, as they directly and indirectly strengthen the self-respect of cultures that predominantly rely on orally transmitted cultures, promote linguistic and cultural diversity, and ensure the sustained access to that kind of documentary heritage. Additionally, they secure jobs and increase the competence of the institutes. Under the prevailing financial conditions of the countries, sustainable audiovisual preservation can only be achieved by cooperation. The projects outlined could meet the high expectations to serve as an example for the cooperative solution of similar problems. Cooperations to manage and further develop all audiovisual holdings of the regions are ongoing, and the competence acquired in the training and practical implementation of the projects can be maintained and further developed for the benefit of all existing audiovisual holdings of the regions. This creation of competence centers also diminishes the risk of declining preservation budgets, and hopefully such projects will be supported even beyond present financial limits.

Further cooperation and not least a cordial friendship connects our institutes and shows that no archive should stay an island in the ocean of digitisation!

Acknowledgements

I would like to thank Dietrich Schüller and Andreas Matzke for their contributions to this article.

UNESCO/UBC
VANCOUVER DECLARATION

UNESCO/UBC VANCOUVER DECLARATION*

The Memory of the World in the Digital Age: Digitization and Preservation

26 to 28 September 2012
Vancouver, British Columbia, Canada

Digital technology offers unprecedented means to transmit and store information. Documents and data in digital form are important for science, education, culture and economic and social development, but assuring their continuity over time is a far from resolved problem. While countries differ greatly as to the possibilities they have to implement policies to address sustainability access to digital resources, the fundamental challenges are universal. Closer collaboration in managing these resources will be beneficial for all.

At present, digital information is being lost because its value is underestimated, because of the absence of legal and institutional frameworks or because custodians lack knowledge, skills and funding. In order to explore these issues in depth and obtain solutions, UNESCO's Director-General convened an international conference: *The Memory of the World in the Digital Age: Digitization and Preservation* from 26 to 28 September 2012 in Vancouver, British Columbia, Canada.

More than 500 participants from 110 countries discussed the key factors affecting the two major aspects of records, documents and data in the digital environment:

- issues pertaining to the digitization of analogue material, and
- issues pertaining to continuity, access, and preservation of authentic, reliable, and accurate digital materials.

As a result of these discussions, the participants agreed that:

1. as enshrined in Article 19 of the Universal Declaration on Human Rights, each individual has the right to seek, receive and impart information through any media and regardless of frontiers (article 19). Citizens exercise this right when they access information in digital form. Trustworthiness and integrity of documentary heritage and documentary systems are therefore a prerequisite for the continued exercise of this right;
2. for analogue documents, digitization can protect valuable documents from deterioration by reducing handling. In the case of audiovisual documents, digitization is the only means of ensuring their survival;
3. many objects are born digital, but without due consideration of the means of ensuring their continuing accessibility, and authentic, reliable, and accurate preservation through time and technological change. These issues of access and preservation apply also to digitized materials;

* www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/unesco_abc_vancouver_declaration_en.pdf

4. a better understanding of the digital environment is essential for the establishment of digital preservation models that respect fundamental legal principles enshrined in institutional regulatory frameworks, and balance access with privacy, right to knowledge with economic rights, and respect ownership and control of indigenous cultural heritage and traditional knowledge in digital format;
5. digital preservation should be a development priority, and investments in infrastructure are essential to ensure trustworthiness of preserved digital records as well as their long-term accessibility and usability;
6. education and training programmes for information professionals must be developed and provided to prepare or reposition them to implement both digitization and preservation practices relevant to the needs of governments and their citizens;
7. there is a pressing need to establish a roadmap proposing solutions, agreements and policies, that ensure long term access and trustworthy preservation. This roadmap should address issues like open government, open data, open access and electronic government. It should dovetail with national and international priorities and be in full agreement with human rights.

Recommendations

Taking current and emerging challenges into consideration, the participants:

Urge the UNESCO secretariat to:

- a. play an active advocacy role to make digital preservation frameworks and practices a reality, by promoting digital objects management and preservation in all appropriate forms, including working with other UN agencies, funds and programmes;
- b. support the work of the international archival, library and museum community to secure an international legal framework of copyright exceptions and limitations to ensure preservation of and access to cultural heritage in digital format, and acquisition of and access to that heritage in a culturally appropriate manner;
- c. collaborate with international professional associations and other international bodies to develop academic curricula for digitization and digital preservation, and implement training programmes and global educational approaches that enhance the capabilities of archives, library, and museum personnel to manage and preserve digital information;
- d. establish a multi-stakeholder forum for the discussion of standardization in digitization and digital preservation practices, including the establishment of digital format registries;

- e. in cooperation with international professional associations and research projects teams, design and publish guidelines, policies and procedures as well as best-practice models in digitization and digital preservation;
- f. support the belief that good management of trustworthy digital information is fundamental to sustainable development by developing and implementing a global digital roadmap under the auspices of the Memory of the World Programme to encourage all relevant stakeholders, in particular governments and the industry, to invest in trustworthy digital infrastructure and digital preservation;
- g. create an emergency programme aiming at preservation of documentary materials endangered by natural disasters or armed conflicts, as well as a programme for the recovery of analogue and digital heritage that is under threat of becoming, or is already, inaccessible because of obsolete hardware and software;
- h. encourage engagement of cultural heritage professionals knowledgeable about digital forensics concepts, methods and tools in order to ensure capture and reliable preservation of authentic, contextualized and meaningful information, and appropriate mediation of access to the information;
- i. update the implementation guidelines of the 2003 UNESCO Charter on preservation of digital heritage and give consideration to the inclusion of preservation of and access to digitized cultural heritage in the proposed recommendation on documentary heritage being examined by the 190th session of UNESCO's Executive Board;
- j. work with national and international research and heritage bodies to develop criteria for assessing whether repositories are, or can be improved to be, trustworthy in terms of their ability to preserve digital holdings;
- k. promote cooperation with international standards bodies in order to increase consistency among different reference sources on digital preservation, and support the development of standards compliant with the principles endorsed by UNESCO.

Urge UNESCO's Member States to:

- a. develop and enforce laws that ensure rights of all citizens to relevant knowledge;
- b. develop public policies enabling and supporting preservation of digital heritage in a rapidly changing technological environment;
- c. promote cooperation between their legislative bodies and archives, libraries and museums and other relevant organizations, in order to develop legal frameworks that support preservation of, and access to, digital cultural heritage;
- d. develop strategies for open government and open data that address the need to create and maintain trust and reliance in digital government records;

- e. provide legal guarantees that information to which citizens are legally entitled be available in an open format;
- f. encourage private sector organizations to invest in trustworthy digital infrastructure and digital preservation;
- g. develop a Recommendation for the promotion of legal deposit laws for digital formats;
- h. establish appropriate oversight body(ies), e.g., Information Ombudsman, to monitor and protect the necessary degree of independence required by archives, libraries, museums and other heritage organizations to preserve and provide access to digital information in such a way that sustains public trust in what information is selected for preservation and how it is preserved;
- i. identify and propose registration of digital documentary heritage on a Memory of the World Register;
- j. ensure that analogue contents will be made available in digital form, to avoid their future neglect in a world of predominant digital information retrieval;
- k. raise public awareness of relevance of digital preservation for the endurance of our cultural heritage;
- l. promote the use of standards and widely recognized guidelines and best practices on digitization and digital preservation among the relevant national organizations and communities.

Urge professional organizations in the cultural heritage sector to:

- a. cooperate with other professional associations, international and regional organizations and commercial enterprises to ensure that significant born-digital materials are preserved by promoting and advocating for digital legal deposit laws;
- b. assist in the development of a cohesive, conceptual and practical vision for a digital strategy capable of addressing the management and preservation of recorded information in all its forms in the digital environment;
- c. encourage their members to take into consideration the reliability, authenticity, copyright ownership and future use of digital information, and to develop policies for all aspects of management and preservation of digital materials;
- d. cooperate with the private sector for the development of products that facilitate the long-term retention and preservation of information recorded in a digital format;

- e. encourage members to identify and evaluate the specific threats to which their digital information is vulnerable, and implement appropriate processes and policies to mitigate these threats.

Urge private sector organizations to:

- a. cooperate with archives, library, museum and other relevant organizations to ensure long-term accessibility to digital information;
- b. adhere to recognized metadata standards designed in cooperation with information professionals for description and/or management of digital resources, in order to enable interoperability of sources that can be presumed authentic and guaranteed reliable and accurate;
- c. take digital preservation issues into consideration when participating in national and international standards initiatives and in their work on multi-jurisdictional and other partnership initiatives where information generated in a digital format is to be retained through the long term.

Thank you to our sponsors!

Diamond Sponsors



Platinum Sponsors



Ministry of Communications
and Information Technologies,
Republic of Azerbaijan



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada



a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA



UNIVERSITY OF TORONTO
LIBRARIES

Gold Sponsors



McGill



InterPARES Project

International Research on Permanent Authentic Records in Electronic Systems

Silver Sponsors



Ministry of Education, Culture and
Science

OCUL Ontario Council of
University Libraries



**UNIVERSITY OF ALBERTA
LIBRARIES**



uOttawa
Library

Supporters



CANADIAN COMMISSION FOR UNESCO
COMMISSION CANADIENNE POUR L'UNESCO
www.unesco.ca



Québec



École des sciences de l'information
School of Information Studies