Preservation and Curation in Institutional Repositories

Alex Ball
Digital Curation Centre, UKOLN, University of Bath

DCC STATE OF THE ART REPORT
Deliverable B 5.2

Version: 1.3
Status: Final
Date: 10th March 2010

Copyright

Catalogue Entry

Title	Preservation and Curation in Institutional Repositories
Creator	Alex Ball (author)
Contributor	Maureen Pennock; Michael Day
Subject	Open Archival Information System (OAIS); repository architectures; preservation planning; preservation metadata; representation information; persistent identifiers; repository certification; repository audit; costing
Description	Institutional repositories were originally intended as a way of giving immediate and wide access to research papers. They are increasingly taking on a role as curators of institutional digital output, requiring the adoption of specific policies and tools for preservation and curation. In the ten years since the first dedicated institutional repository software was released, a range of tools have been developed to assist in everything from drawing up preservation plans and policies to extracting preservation metadata from files, alongside modular architectures for linking all the tools together. While the uptake of these technologies and techniques in repositories is modest, there are encouraging signs of progress.
Publisher	University of Edinburgh; UKOLN, University of Bath; HATII, University of Glasgow; Science and Technology Facilities Council
Date	27th February 2009 (creation)
Type	Text
Format	Portable Document Format version 1.4
Language	English
Rights	© 2010 Digital Curation Centre, UKOLN, University of Bath

Citation Guidelines

Alex Ball. (2010). *Preservation and Curation in Institutional Repositories* (version 1.3). Edinburgh, UK: Digital Curation Centre.

Contents

1 Introduction

The concept of online repositories of scientific publication has been around as long as the World Wide Web. The first e-print repository was arXiv, established in August 1991 at the Los Alamos National Laboratory and initially serving high energy theoretical physics.[1] Physicists in this area already had a culture of sending each other hard-copy pre-prints of articles as they were completed, as a more rapid form of dissemination than journals could provide. The `hep-th` database provided them with a dissemination route that was cheaper and much easier to administer, as well as being even faster than the paper-based systems (Ginsparg, 1994). The arXiv repository has since expanded to cover most other areas of physics, as well as areas of mathematics and computer science.

The success of arXiv was followed by the launch of similar services for other disciplines and large institutions, and eventually lead to the formation of the Open Archives Initiative (OAI) (Van de Sompel & Lagoze, 2000). One of the outcomes of the first meeting of the OAI was the adaptation of the software underlying CogPrints to make it easier for institutions to set up their own repositories: this software was named EPrints and released in 2000.[2] Since then, other institutional repository systems have emerged, notably DSpace and Fedora.[3]

The question of whether institutional repositories should have preservation responsibilities was raised early on (Pinfield & James, 2003). In some quarters, there was (and still is) strong resistance to the notion, the argument being that repositories are solely for the purposes of accelerating the dissemination and widening the impact of high quality research; preservation should more properly target the 'official' printed record, published in journals and held by libraries. There are, of course, arguments counter to this view. One is that the usefulness of e-prints does not cease once the printed versions are published. Open access to e-prints means that researchers can still access the material even if they do not belong to an institution that subscribes to the journal, or one that could obtain a copy through interlibrary loan. Another argument is that institutional repositories can hold more than just surrogates of journal articles: for example, expanded versions of articles, underlying data, unpublished conference papers, teaching and learning resources, multimedia presentations, or corporate material (administrative records, publicity materials, etc.). Thus, while research and development in the areas of preservation and curation in the context of institutional repositories was slow to begin with, it has accelerated and is gaining more mainstream attention. It is telling, for example, that the JISC created a combined Repositories and Preservation Programme in

1. Main arXiv Web site, URL: `http://arxiv.org/`
2. CogPrints Web site, URL: `http://cogprints.org/`; EPrints Web site, URL: `http://www.eprints.org/`
3. DSPace Web site, URL: `http://www.dspace.org/`; Fedora Commons Web site, URL: `http://www.fedora-commons.org/`

April 2006 to further the work of both its Digital Repository Programme and its Digital Preservation and Asset Management Programme.

This report provides a snapshot of the state of the art of preservation and curation in an institutional repository context, noting areas of recent and current research and development. It should be of interest principally to institutional repository managers and others concerned with the strategic planning for these services. The report begins with a brief introduction to preservation and curation, followed in chapter 3 by a summary of the current provision for these activities in EPrints, DSpace and Fedora. Some repository models and architectures relevant to preservation and curation are presented in chapter 4 and chapter 5 respectively, while a selection of preservation planning tools of possible use in a repository context are described in chapter 6. Pertinent developments in metadata are reviewed in chapter 7, while tools for working with such metadata are presented in chapter 8. Technologies that assist in performing emulation, reverse engineering and migration are described in chapter 9. The issue of identifiers for repository materials is tackled in chapter 10. A selection of guidelines and tools for auditing curatorial aspects of institutional repositories is presented in chapter 11, and a selection of tools for calculating the costs and benefits of curation is presented in chapter 12. Finally, some conclusions are drawn in chapter 13.

2 Preservation and Curation

The term 'digital curation' is relatively new, having been coined in 2001 as the title for a seminar on digital archives, libraries and e-Science (Beagrie, 2006) It was given perhaps its most precise formulation by Lord and Macdonald (2003, p. 12), who proposed the following definitions for curation, archiving and preservation.

> **Curation.** The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials.
>
> **Archiving.** A curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.
>
> **Preservation.** An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.

Note that while the definitions are for curation, archiving and preservation *simpliciter*, the authors have digital objects, and specifically datasets, in mind. While these definitions have been influential in the field of digital curation,[1] there are other, looser definitions in circulation. For example, the Digital Curation Centre (DCC) defines digital curation as 'broadly interpreted, . . . about maintaining and adding value to a trusted body of digital information for current and future use', with curation in general as 'the active management and appraisal of data over the life-cycle of scholarly and scientific interest' (Digital Curation Centre [DCC], 2007).

The DCC has produced a model for aligning curation tasks with the lifecycle stages of a digital object, intended as a planning tool for data creators, curators and users (Higgins, 2008). A graphical representation of the model can be seen as Figure 2.1.

At the centre of the Model are the digital data, which are here identified with simple and complex digital objects or databases. The Model notes three levels of full-lifecycle actions:

- *Description* and (management of) *Representation Information*. The creation, collection, preservation and maintenance of sufficient metadata (and recursions thereof) to enable the data to be used and re-used for as long as they have value to justify continued curation.

1. See, for example, Hitchcock, Brody, Hey and Carr (2005), Bose and Reitsma (2006), Livingston and Nastasie (2007), Law, Peng and Demian (2005), and Joint Information Systems Committee [JISC] (2003).

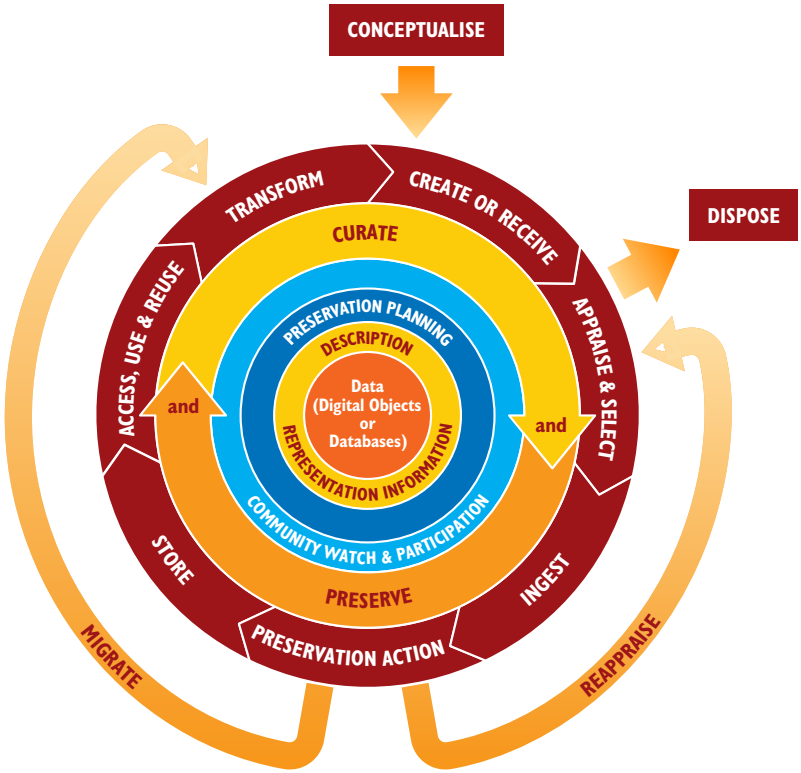

Figure 2.1: DCC Curation Lifecycle Model

- *Preservation Planning*. Strategies, policies and procedures for all curation actions.
- *Community Watch and Participation*. The observation of the target community of the data, in order to track changes in their requirements for the data, and participation in the development of standards, tools and software relevant for the data.

The fourth level, *Curate and Preserve*, properly describes most of the actions in the model, but is used here to represent the execution of the planned management and administrative actions supporting curation.

The sequential actions are not exclusively concerned with curation, but rather represent stages of the data lifecycle which ought to have a curation component. They begin with *Conceptualise*: the planning stages of the data generation and collection activities. Aspects such as the capture method will be informed by considerations other than curation – the scientific rigour of the method will be particularly important – but matters such as how the data will be stored, what budget to allocate for curation, and how collection of information important for curation may be automated or otherwise simplified, should be dealt with at this stage.

The curation lifecycle proper begins with the *Create or Receive* stage, where 'create' refers to original data generated and recorded by researchers, and 'receive' refers to pre-existing data collected from other sources. The curation activities at this stage centre on ensuring that all data is accompanied by sufficient administrative, descriptive, structural and technical metadata; ideally, pre-existing data should have these already,

but as different researchers and repositories inevitably work to different standards they should be checked for consistency with local policies.

In the next stage, *Appraise and Select*, researchers or data specialists evaluate and select the data to keep for the long term according to documented guidance, policies or legal requirements. Some data may be sent for *Disposal*: this may involve transferring the data to another custodian, although it could mean simple or secure destruction. Again, the nature of the disposal should be driven by documented guidance, policies or legal requirements. The remaining data are sent for *Ingest* by the normal custodian, be that an archive, repository, data centre or some other service. The Ingest stage immediately leads on to the *Preservation Action* stage, which involves an array of different activities: quality control, cataloguing, classifying, generating fixity data, registering semantic and structural metadata, and so on. Any data that fail quality control checks are returned to the originator for further appraisal. This should result either in improvements in the quality of the data (e.g. corrections to data transfer procedures, improved metadata, repackaging of data) and reselection, or disposal. Some data may need to be migrated to a different format, either to normalise it within the system or to reduce risks arising from hardware or obsolescence.

Once the data have completed the *Preservation Action* stage, they pass into *Storage*. This principally refers to the initial committal of the data to storage, but various long-term actions that ensure data remain secure may also be associated with this stage: maintaining the storage hardware, refreshing the media, making backup copies, checking for fixity, and so on.

Once the data have been safely stored, they enter a period of *Access, Use and Re-use*. Curation actions associated with this stage are focussed on keeping the data discoverable and accessible to designated users and re-users. This includes, for example, surfacing descriptive metadata through custom search interfaces or public APIs, and ensuring the preservation metadata held for the data continue to meet the requirements of the designated users and re-users.

Aside from ongoing preservation activities, the story of the archived data considered as an object stops at that point, but several events may cause progression to the *Transform* stage of the lifecycle. A key piece of software or hardware may approach obsolescence, therefore triggering an action to migrate the data to a new format, or a (re-)user may request a subset or other derivation of the data. The end result is a new set of data which starts the lifecycle again. Data created for a repository's internal purposes will, by its nature, pass through the early lifecycle stages rapidly, while data supplied to a (re-)user will progress as normal.

3 Repository software

According to the Directory of Open Access Repositories, the two most popular repository software systems in use are DSpace and EPrints; figures from the Fedora Commons suggest Fedora is the third most popular (OpenDOAR, 2009; Smith, Davis & Staples, 2008). All three systems have some support for preservation activities as standard, though the extent to which this support is used in a particular instance will vary according to the local policies, procedures and system configuration.

3.1 EPrints

EPrints is developed by the University of Southampton, and has its origins in the software underlying the CogPrints subject repository. The version current at the time of writing is version 3.2, released in March 2010. Version 3 of the software introduced three features aimed at improving support for preservation in EPrints (Brody, Carr & McSweeney, 2010):

- *History module*. This module provides for each repository object a log of changes made to it or the record for it within the repository system. Currently, this is only used to track changes to the record, for audit purposes, but could be used to track preservation actions such as format migrations.
- *METS and DIDL export plugins*. These plugins allow complex objects – that is, objects consisting of more than one file – to be exported as a package in either METS format (METS Editorial Board, 2007) or MPEG-21 DIDL format (ISO/IEC 21000-2, 2005).
- *Creative Commons licensing*. There may in future be some doubt as to whether an institutional repository has the necessary rights and permissions to perform preservation actions on the deposited contents. A licence option has therefore been added to the deposit process, allowing authors to explicitly grant the those permissions to the repository; the software stores these permissions along with the rest of the object metadata.

Further support for preservation in EPrints is planned and will be developed as part of the KeepIt Project (see section 5.7). Among the tools and services proposed are facilities for long-term reliable storage, file and format classification tools, risk analysis tools and format migration tools.[1]

1. EPrints Digital Preservation Web site, URL: `http://preservation.eprints.org/`

3.2 DSpace

DSpace was initially developed as a collaboration between the Massachusetts Institute of Technology and Hewlett-Packard Laboratories, and was first released in 2002. It is now developed under a community development model, with strategic leadership from the DSpace Federation, later the DSpace Foundation, which joined with Fedora Commons to form DuraSpace. The version current at the time of writing is version 1.6, released in March 2010.

DSpace performs bit-level preservation on all deposited objects. In addition, it supports migration or emulation tools for a selection of common, published formats such as TIFF, SGML, XML, AIFF and PDF. Of the formats not supported in this way, the DSpace distinguishes between known and unknown formats: known formats are given their respective format identifiers, while unknown formats are marked as a generic byte-stream using the MIME type 'application/octet-stream'. In preservation terms, known formats are closed but so common that one can be reasonably confident that tools will become available for preserving files in those formats, whereas unknown formats are not common enough to inspire such confidence ('Frequently Asked Questions', n.d., n.d.).

3.3 Fedora

Fedora (Flexible Extensible Digital Object Repository Architecture) began as a project of Cornell University's Digital Libraries Research Group in 1997, though the software was not released until 2003. The name refers both to the abstract architecture and its implementation by the Fedora Repository Project. It is now developed jointly by Cornell University and the University of Virginia, supplemented by additional utilities developed by the Fedora Commons user community; strategic leadership comes from DuraSpace, the organisation formed from joining Fedora Commons and the DSpace Foundation in May 2009. The version of Fedora current at the time of writing is version 3.3, released in December 2009.

Fedora Commons set up a working group looking specifically at preservation issues in 2005; since then, several features have been added to or improved in Fedora the better to support preservation (Bodhmage, 2005'Fedora Repository 3.3 Documentation', 2010; Glick, Wilczek & Dockins, 2006; Jantz et al., 2006).

- *Object versioning.* Fedora supports versioning of both datastreams (data, metadata) and disseminators (rendering services), and preserves the link between a version of the datastream and the corresponding version of the disseminator. The different versions are kept logically within the same digital object and an audit trail of relationships between versions is also in the object's metadata. Versioning is enabled by default, but can be switched off for an entire datastream, or just for specific types of changes.
- *Resource index.* Each digital object can have two datastreams for storing information about relationships: one for the object's relationships to other objects, and

the other for internal relationships between datastreams. These relationships are expressed in RDF using the Fedora Relationships ontology, though other ontologies may be used in addition. A separate resource index is maintained for querying these relationships.

- *Package ingest.* Fedora can ingest objects, both singly and in bulk, that have already been packaged using METS (METS Editorial Board, 2007), Atom (Nottingham & Sayre, 2005) or Fedora's own internal packaging format, FOXML.
- *Mirroring.* It is possible to set up read-only mirrors of Fedora repositories using the multicast journal transport feature.
- *Access control.* It is possible to specify complex access policies for repository content using eXtensible Access Control Markup Language ('OASIS eXtensible Access Control Markup Language (XACML) TC', 2009).
- *Format characterisation.* A third-party module is available for validating file formats and extracting metadata from datastreams using JHOVE ('Other VTLS Open Source Components', 2008).

4 Open Archival Information System Reference Model

The Open Archival Information System (OAIS) Reference Model is a conceptual framework for describing functions, roles and responsibilities of archival repositories (Consultative Committee for Space Data Systems [CCSDS], 2002). While the Model was developed with archives of Space Science data in mind, it is framed sufficiently generally to apply to archives holdings all kinds of digital and physical resources, and indeed has become widely adopted in digital repository and digital curation communities.

The Model has two principal aims: to establish a standard terminology for describing the features of archival repositories, and to establish a minimum level of functionality for archival repositories. The Model does not prescribe that the functions or features of the repository should be implemented in any particular way, only that it should be possible to map between the implemented functions and features of the repository, and those described by the Model.

The OAIS Reference Model contains two detailed models. The first is an Information Model which describes the types of objects and information that an OAIS deals with (see Figure 4.1). The resource to be preserved is the Data Object; the extra information needed to make sense of it is called Representation Information. Representation Information can be structural (e.g. file formats), semantic (e.g. a code book) or some

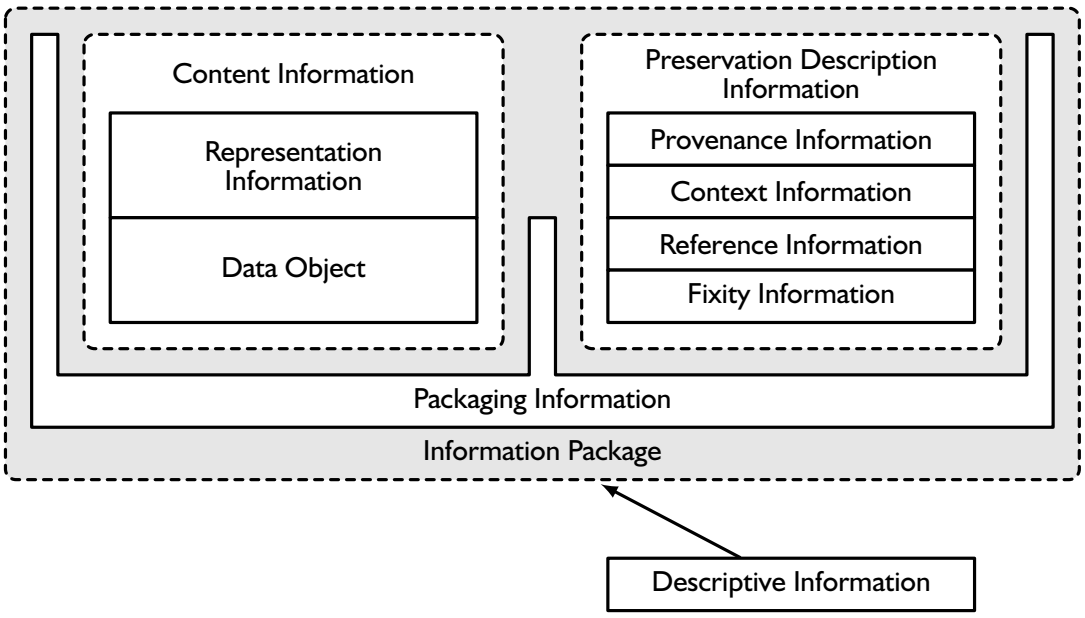

Figure 4.1: OAIS Information Model (CCSDS, 2002, pp. 2-5, 4-37)

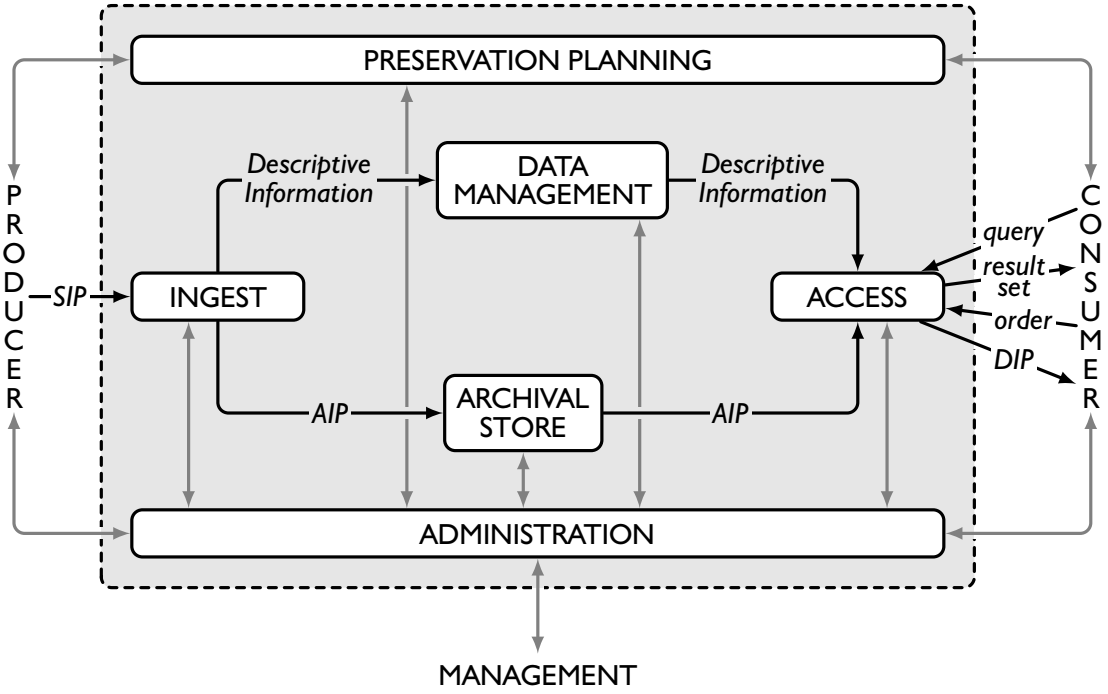

Figure 4.2: OAIS Functional Model (CCSDS, 2002, p. 4·1)

other type (e.g. software); aggregated with a Data Object, it forms Content Information, and combined with it, it forms an Information Object. The model identifies four further types of information as important for preservation purposes: Provenance Information (the source and processing history of the Data Object), Context Information (how the Data Object relates to other Data Objects), Reference Information (e.g. identifiers) and Fixity Information (e.g. checksums). These pieces of information are related together using Packaging Information (e.g. a manifest) to form an Information Package. Further Descriptive Information (e.g. a catalogue record) should be used to enable the package to be found in a retrieval system.

The second detailed model is the Functional Model, describing the functions, processes, roles and responsibilities of an OAIS (see Figure 4.2). While the Model goes into some depth about the processes associated with each functional entity, they may be summarised as follows. Producers submit data and associated metadata to the OAIS in the form of a Submission Information Package (SIP). The *Ingest* entity of the OAIS processes the SIP – performing quality assurance and normalizing the metadata – to produce an Archival Information Package (AIP) and accompanying Descriptive Information. The AIP is transferred to the *Archival Storage* entity, which also performs backups, media refreshment and so on. The Descriptive Information is transferred to the *Data Management* entity, which integrates the information into a catalogue database. When Consumers wish to retrieve a resource from the OAIS, they do so via the *Access* entity of the OAIS. Initially, the Consumer issues a query to the Access entity, which passes it to the Data Management entity. Data Management performs the query and returns a result set, which the Access entity passes back to the Consumer. Alternatively, the Consumer may issue a report request, requiring several queries to be performed; the report is also generated by the Data Management entity before being passed back through the

Access entity. Finally, the Consumer places an order through the Access entity; the Access entity retrieves both the Descriptive Information from the Data Management entity and the AIP from the Archival Storage entity, and combines them into a suitable Dissemination Information Package (DIP). The Access entity then delivers the DIP to the Consumer.

The *Administration* entity establishes policies and procedures, negotiates submission agreements, performs audits and manages system configuration. The *Preservation Planning* entity develops preservation strategies, migration plans and packaging designs, and monitors both the community served by the OAIS and the wider technological environment for changes that would affect requirements for Representation Information or Preservation Description Information. Omitted from the diagram are the *Common Services* that underlie the operation of the OAIS: operating system services, network services and security services.

The OAIS Reference Model has proved highly influential, not only with the large data centres for which it was intended, but also for institutional repositories. Allinson (2006) gave a cautious welcome to using OAIS in repositories, as a tool for ensuring good practice and encouraging a focus on preservation. The developers of aDORe, DSpace and Fedora have embraced OAIS terminology to a greater or lesser extent (Bass et al., 2002; Celeste & Branschofsky, 2002; Van de Sompel, Bekaert, Liu, Balakireva & Schwander, 2005; 'Packager Plugins', 2007; 'Asset Store', 2009; Fedora Development Team, 2005; 'Fedora Service Framework', 2007), and it also features prominently in repository projects such as RODA (Ramalho et al., 2008), and the repository architecture models produced by CASPAR, SHAMAN, PRESERV, and SHERPA DP (see subsections 5.4, 5.6, 5.7 and 5.8 respectively).

5 Preservation architectures for repositories

While the OAIS Reference Model has been influential as an abstract model of repository activity, it does not recommend any particular architecture for repositories. Such architectures and the software populating them must be designed separately. As noted above, institutional repository software was initially designed with rapid, open access in mind rather than preservation, so the core architecture of such software typically lacks features mentioned or implied by OAIS. There have therefore been several attempts to design and implement comprehensive preservation architectures that can be merged with existing repository architectures. The examples of this presented below are PANIC, Seamless Flow, the CRiB, the CASPAR Preservation Workflow, the Planets Interoperability Framework and the SHAMAN application solution environments.

Other projects have taken a different approach, looking at how curation and preservation might be served in the institutional repository's relationships with other repositories and systems. SHERPA DP considered how institutional repositories might use a third party preservation service running a dark archive repository. LOCKSS and CLOCKSS provide a mechanism for institutional repositories to form a network for mutually backing-up shared content. RepoMMan and REMAP took the opposite perspective and investigated how institutional repositories could act as a preservation service for other institutional systems. Finally, DRIVER developed the means to harmonise many different repositories and integrate their contents.

Bridging these two approaches is the PRESERV Project, which considered both the internal and external methods of adding preservation capability to a repository. As well as developing architectures similar to those of SHERPA DP, it was also responsible for adding additional preservation capabilities to institutional repository software.

5.1 PANIC

The PANIC (Preservation web services Architecture for New media, Interactive Collections and scientific data) Project was undertaken by the Distributed Systems Technology Centre (DTSC) and the University of Queensland between 2003 and 2006.[1] The aim of the Project was to provide an architecture that could be added to repository systems to allow them to support the preservation of digital objects. The PANIC system architecture (see Figure 5.1) is based around three key processes: preservation metadata capture, obsolescence detection and notification, and preservation service discovery

1. PANIC Project Web site, URL: http://www.itee.uq.edu.au/~eresearch/projects/panic/

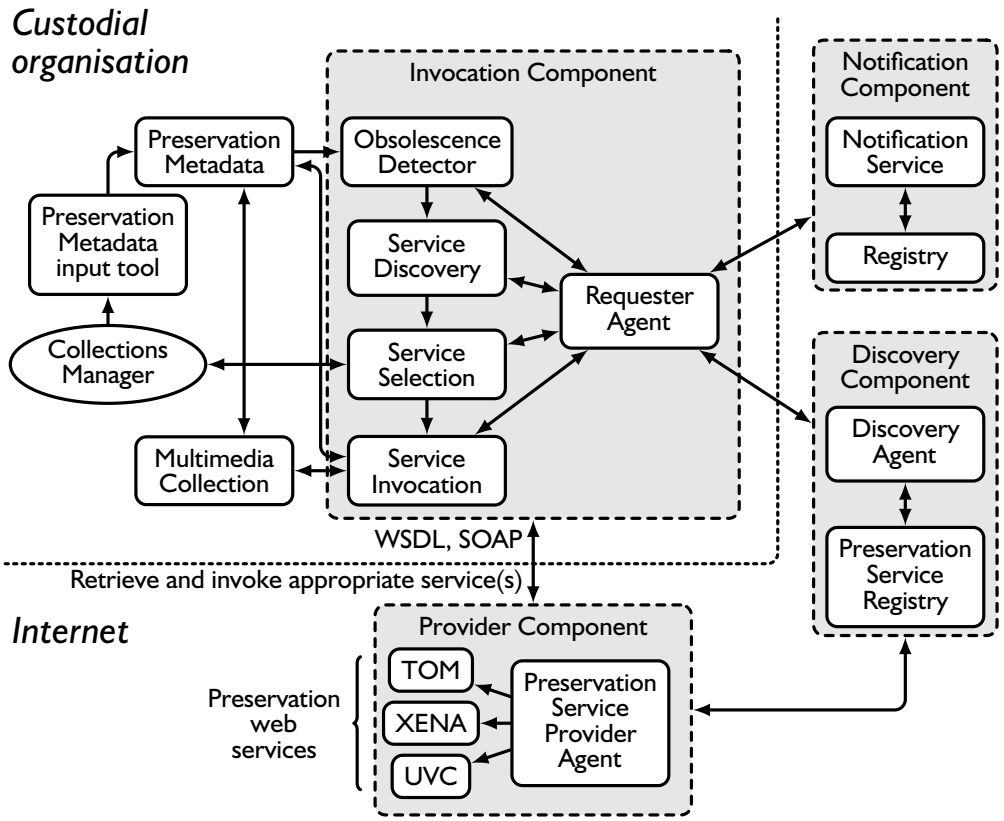

Figure 5.1: PANIC system architecture (Hunter & Choudhury, 2006, p. 176).

and invocation. To support the capture of preservation metadata, the Project developed PREMINT (Preservation Metadata Input Tool), which provides a user-friendly interface for writing preservation metadata into both METS and MPEG-21 DIDL packages (see section 8.5). It also encoded PREMIS entities as OWL ontologies ([W3C], 2004). To support obsolescence detection and notification, the Project developed three databases – a software version registry, a format registry and a recommended format registry – and associated metadata profiles as a proof of concept, although it recommended using third party Web services as they became available. To support preservation services, the Project developed extensions to the OWL-S Web service ontology (Martin et al., 2004) specifically for preservation services, profiles and processes, and also a Discovery Agent called *Semantic Matchmaker* (Hunter & Choudhury, 2006).

5.2 CRiB

The CRiB (Conversion and Recommendation of Digital Object Formats) is a service-orientated architecture developed at the University of Minho, Portugal.[2] It is aimed at cultural heritage institutions and, as its name suggests, uses migration as its principal method for maintaining access to resources. The CRiB architecture contains services for identifying the format of a digital object, monitoring for near-obsolete formats,

2. CRiB project page, URL: `http://crib.dsi.uminho.pt/`

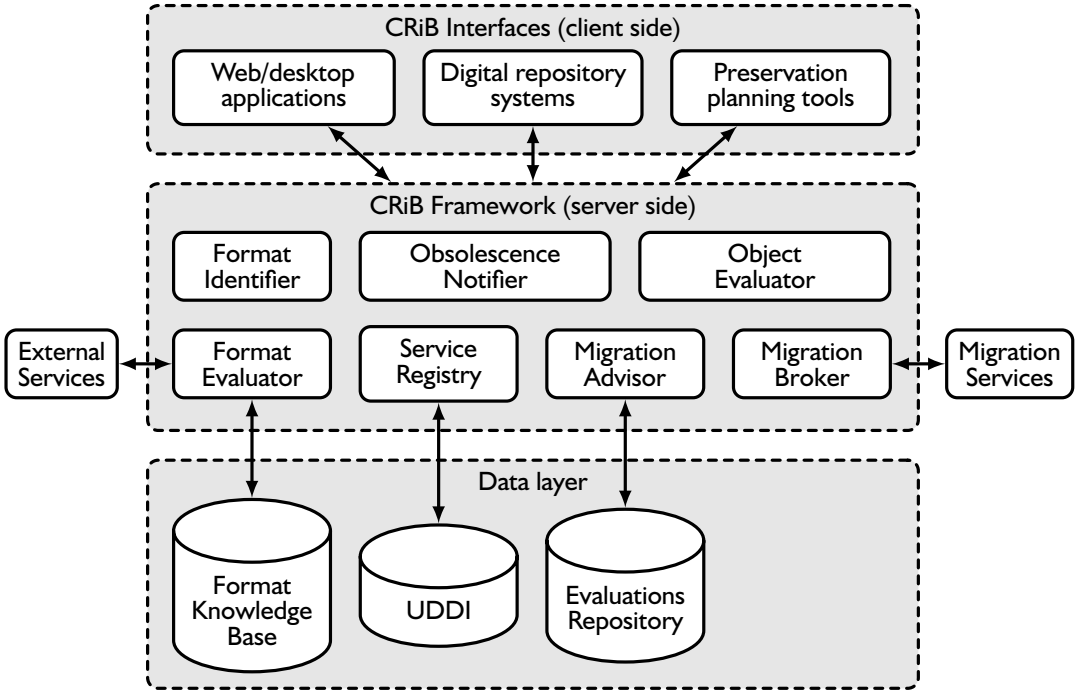

Figure 5.2: CRiB system architecture (Ferreira, Baptista & Ramalho, 2009; Ramalho et al., 2008).

comparing the capabilities of different formats, registering available migration services, recommending migration pathways, performing migrations and evaluating the outcomes of migrations (see Figure 5.2). Each of these services may be adjusted to suit a particular institution's needs. It has interfaces both for desktop/Web-based clients and digital repository systems such as DSpace, EPrints and Fedora.

While still under development, the CRiB has reached sufficient maturity to form part of the Portuguese National Archives' Repository of Authentic Digital Objects (RODA), a multimedia repository of outputs from national public institutions (Ramalho et al., 2008). The development team are also considering other ways in which the CRiB may be integrated with other systems; for example, Becker et al. (2008) describe how the CRiB may be used in conjunction with a preservation planning service such as PLATO (see section 6.1). Evaluation criteria for determining the success or otherwise of migrations have been drawn up for conversion processes in general, and for still images and text documents in particular. Technical criteria for comparing still-image formats have also been compiled.

5.3 Seamless Flow

The (UK) National Archives (TNA) Seamless Flow Programme ran between 2005 and 2008, setting up an active preservation framework for governmental electronic records (Brown, 2007). The framework is based on three key activities: characterisation, preservation planning and preservation action.

When an object is ingested into the digital repository, the DROID tool (see section 8.2) is used to identify the format. The framework looks up the format in the PRONOM registry (see section 7.1 and discovers appropriate tools for checking if the object is well-formed (syntactically compliant with the format specifications) or valid (well-formed and compliant with semantic constraints); these tools are deployed automatically. Finally, the JHOVE tool (see section 8.1) is used to extract preservation metadata from the object and generate information about its significant properties.

The objects within the repository are regularly checked for risk factors using the PRONOM registry – risks relating to the number of supporting software tools, the openness of the format, compression techniques used, and so on. If at any point the risk factor for an object or a class of objects rises above a critical threshold, the framework interrogates the PRONOM registry to determine appropriate alternative formats for the at-risk objects, and possible migration pathways between the current and preferred formats. These are presented to an archivist for testing and review. Once a pathway (preservation plan) has been certified as suitable for this situation, with regard to the preservation of defined significant properties, the framework goes ahead and carries out the plan. Following the execution of the preservation plan, the newly-migrated files are characterised and checked against the characterisations of the originals to ensure that none of the significant properties have been damaged.

5.4 CASPAR

CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval) was an EU-funded integrated project looking at a range of preservation issues from across the curation lifecycle and across diverse disciplines (Giaretta, 2007, 2009).[3] The project ran from April 2006 to September 2009, and in that time developed an architecture and workflow for preservation, along with exemplar tools fulfilling the roles in the architecture (see Figure 5.3). These tools include: REPINF (Representation Information Toolkit), VIRT (Virtulisation), REG (Registry), PACK (Packaging), PDS (Preservation Data Stores), FIND (Finding Aid), KM (Knowledge Manager), POM (Preservation Orchestration Manager), DAMS (Data Access Manager and Security), DRM (Digital Rights Manager) and AUTH (Authenticity) (CASPAR Consortium, 2007).

The project set up three preservation testbeds in which the CASPAR Preservation Workflow was applied to different disciplines: Earth observation satellite data, electronic music and models, photographs, etc. from UNESCO Cultural Heritage Sites. These testbeds had several aims: to ensure that the CASPAR Preservation Workflow could be used successfully in diverse contexts and with diverse data types; to determine the Representation Information and Preservation Description Information needed to support the continuing use of these diverse data types over the long term; and to investigate the how different preservation techniques – migration, emulation, dark archiving of source code, hardware preservation, reconstructing software from documentation – suit different circumstances. The testbeds also provided an opportunity to produce tools

3. CASPAR project Web site, URL: `http://www.casparpreserves.eu/`

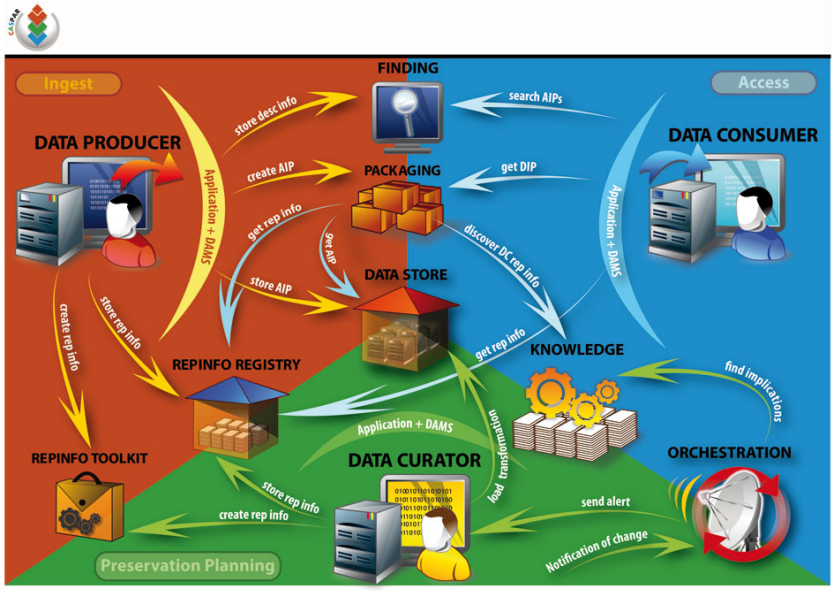

Figure 5.3: The CASPAR Preservation Workflow (Giaretta, 2009, p. 180)

to assist practitioners in collecting preservation metadata and producing Submission Information Packages (CASPAR Consortium, 2009; Šimko, Máša & Giaretta, 2009).

5.5 Planets

Planets (Preservation and Long-term Access through NETworked Services) is an EU-funded project to build practical curation services and tools for cultural and scientific assets.[4] It is co-ordinated by the British Library and involves national libraries, national archives, universities, university libraries and commercial organisations from across Europe. The preservation architecture developed by the project is known as the Planets Interoperability Framework (see Figure 5.4). This framework will be fully implemented, and is designed to run in parallel with existing repository software, ingesting material direct from the repository and making the preservation outputs available for ingest back into the repository.

At a high level, the architecture centres on a gateway server (roughly corresponding to the 'core elements' in the figure). The gateway has two sets of interfaces: Portal Services, which provide a secure means for users to interact with the server, and the WFlow API, which allows the server to interact with preservation services as part of a preservation workflow (Schmidt et al., 2009). Over 50 services have been adapted for use with Planets Interoperability Framework, covering actions such as the format-specific characterisation of files, comparison of the metadata or properties of related files, identifying a file's format, migrating files from one format to another, editing files (e.g. enlarging or cropping images), validating a file's compliance with its format specifications, and rendering a file for end-user consumption. The principal

4. Planets project Web site, URL: `http://www.planets-project.eu/`

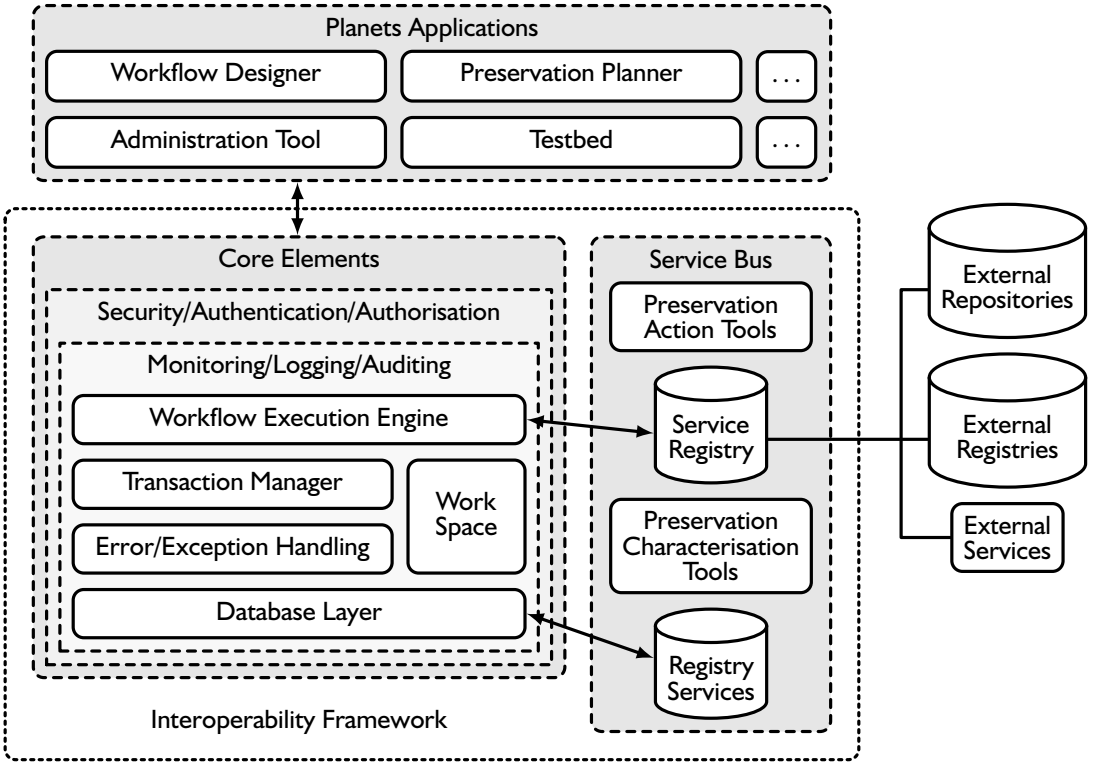

Figure 5.4: The Planets Interoperability Framework (adapted from King, 2008, p. 8).

database components underlying the gateway server are the Core Registry (containing representation information and significant properties information about digital object types and formats, along with characterisation and preservation action tools), the Service Registry (containing information about Planets-compatible services) and the Data Registry (containing the digital objects acted on by the system). As well as graphical user interfaces to core elements of the framework, Planets has also produced task-based user-facing tools, notably the Planets Testbed for evaluating preservation strategies, and Plato, a decision support tool for preservation planning (see section 6.1). Planets started in June 2006 and will run for four years; following this, it is hoped that the work of the project will be carried forward by a not-for-profit membership organisation (Farquhar, 2009).

5.6 SHAMAN

SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) is an EU-funded project to lay the conceptual and technical groundwork for a Grid-based, distributed digital preservation framework.[5] Among the deliverables for this project will be a set of core services and reference implementations of preservation tools. The services will include the integrating Data Grid, a Digital Library, a Persistent Archive, a Context Representation, Annotation and Preservation service, a Deep Linguistic Analysis service

5. SHAMAN project Web site, URL: `http://shaman-ip.eu/`

and Semantic Representation and Annotation technologies. Along with a selection of preservation tools, these services will be integrated into application solution environments and tested in three different domains: cultural memory institutions, industrial design and engineering, and e-Science. SHAMAN began work in 2008 and will run for three years (Innocenti et al., 2009; Watry, 2007).

5.7 PRESERV and KeepIt

The PRESERV Project was a collaboration between the Universities of Oxford and Southampton, and The National Archives (TNA); it ran in two phases between February 2005 and January 2007, and July 2007 and March 2009 respectively (Hitchcock, Hey, Brody & Carr, 2007; Hitchcock, Tarrant & Carr, 2009).[6] The original aim of the Project was to develop an OAIS Ingest service for EPrints-based repositories that linked to TNA's PRONOM service for identifying and verifying file formats. In the course of completing this work, the Project shifted focus to providing more generic repository harvesting interfaces interacting with a suite of lightweight, Web-based preservation services.

The Project performed a survey of repository policies, finding that while preservation policies were rare, and preservation actions were largely restricted to byte preservation, many repositories have policies on submission formats and most transform ingested materials to a different format (usually PDF) (Hitchcock, Brody, Hey & Carr, 2007c).

Among the outcomes of PRESERV were a set of three models for linking preservation services with institutional repositories, expressed in terms of the OAIS Reference Model (Hitchcock, Brody, Hey & Carr, 2007a). The *service provider* model has the institutional repository providing the Data Management function and mediating information package orders from the Access entity, while a preservation service provider undertakes to provide the Archival Storage entity. In the *institutional model*, the institution uses its institutional repository to perform Data Management and rudimentary Archival Storage functions, with the remaining Archival Storage functions performed by an in-house preservation repository. The third, *institutional repository* model envisions an institutional repository with built-in preservation support fully playing the roles of both the Data Management and Archival Storage entities. Hitchcock et al. estimate that the costs associated with these models would run high to low respectively.

The Project's work on integrating preservation services into repositories led to two innovations. The first was the inclusion of an audit history module added to EPrints v3.0, allowing the processing history of ingested material to be recorded and tracked. The second, arising from work on applying PRONOM to two pilot repositories, was an idea for an OAI-PMH-based Web service for producing file format profiles for many different repositories at once. This was achieved in a spin-off project, details of which may be found at section 6.3.

Although PRESERV 2 has now finished, the strategies, policies and services the project outlined will be developed and put into practice in a series of exemplar repositories as

6. PRESERV Project Web page, URL: `http://preserv.eprints.org/`

part of KeepIt, a JISC-funded project that started in April 2009 (Tarrant & Hitchcock, 2009).

5.8 SHERPA DP

The initial SHERPA DP Project investigated the provision of preservation services for e-print archives operated by the members of the SHERPA consortium.[7] The Project developed a model for providing a preservation service as a dark archive running in parallel to a content provider service (typically an institutional repository): in OAIS terms, the Access function of the content provider feeds into the Ingest entity of the service provider, and vice versa. The Project went on to develop a demonstrator service, creating a framework into which software tools such as Fedora, DROID and JHOVE, and standards such as METS and PREMIS, could be plugged. Further work was performed to develop a business and cost model based on the outputs of the LIFE Project (see section 12.2). The Project also performed a survey of the level of consistency in the metadata held by repositories, and considered what metadata could and should be stored by a preservation service provider (Knight & Anderson, 2007).

SHERPA DP ran for two years from March 2005, and was succeeded in March 2007 by SHERPA DP2 Project.[8] This latter two-year project expanded the preservation service model from a simple dual provider system to a more disaggregated model, allowing for content providers running several interconnected archival stores and interfaces, and for preservation service providers relying on further service providers. The Project also investigated different methods by which preservation service providers could obtain content from their clients, further investigated the preservation service requirements of content providers, and extended the SHERPA DP metadata specifications to cover more content types (Knight, 2009).

5.9 LOCKSS and CLOCKSS

Lots of Copies Keep Stuff Safe (LOCKSS) is a programme led by Stanford University, with the aim of replicating libraries' print-based acquisition paradigm for digital assets.[9] Whereas with printed journals, say, libraries buy and take possession of issues as they are published, with e-journals it is more common for publishers to lease online access to a selection of issues. LOCKSS provides repository software specifically tailored for creating a local cache of e-journal articles, enabling continuing access to those issues (from the same URLs) should the journal subscription be cancelled. The name derives from the preservation strategy used by the repository software: where several repositories hold a copy of the same journal article, they each act as backup for the rest, so that if a copy in one repository degrades, it can be repaired using the other copies. The mechanism by which this occurs is designed so that LOCKSS repositories can only

7. SHERPA DP project page, URL: `http://www.sherpa.ac.uk/projects/sherpadp.html`
8. SHERPA DP2 project Web site, URL: `http://www.sherpadp.org.uk/`
9. LOCKSS Programme Web site, URL: `http://www.lockss.org/`

receive replacement data from other repositories, as opposed to new content (Maniatis, Roussopoulos, Giuli, Rosenthal & Baker, 2005). The software uses migration-on-demand as its principal method of keeping the content accessible in the long term.

The main ('public') network of LOCKSS repositories is co-ordinated by Stanford University, but this does not preclude other organisations setting up their own ('private') network, provided that all members of that network are also members of the LOCKSS Alliance (Reich, 2009). Examples of such private networks include: the Council of Prairie and Pacific University Libraries (COPPUL) Consortium; DataPass, ICPSR, University of Michigan; and LOCKSS Docs (US Federal documents). One particularly notable example is Controlled LOCKSS (CLOCKSS), a collaboration between publishers and research libraries to provide an escrow service for journal content. CLOCKSS acts as a dark archive while the journal content is available from publishers, but once this is no longer the case, open access to the content is provided by one or more CLOCKSS repositories. To date, three discontinued journals have been made available in this way.[10]

5.10 RepoMMan and REMAP

The RepoMMan (Repository Metadata and Management) Project attempted to integrate use of an institutional repository into the workflows of academic and administrative users, focussing on management actions (depositing, accessing, sharing, publishing items) and the metadata needed to support them.[11] It ran from 2005 to 2007 and was succeeded by another two-year project called REMAP (Records Management and Preservation).[12] The latter Project investigated how UK Higher Education Institutions might use a digital repository to support records management and digital preservation. To do so it extended the RepoMMan workflow model to include the full lifecycle of digital objects, and introduced an orchestration tool for working with external preservation services, and a notification layer handling three types of repository alert (event-based, time-lapsed and status) (Green, Awre, Sherratt, Lamb & Dolphin, 2007).

5.11 DRIVER

DRIVER (Digital Repository Infrastructure Vision for European Research) is an EU-funded project aiming to link together digital repositories across Europe with a robust and flexible infrastructure.[13] Phase One of the Project ran from June 2006 to November 2007, and delivered a testbed system aggregating Open Access content from 70 repositories across five European countries (Feijen, Horstmann, Manghi, Robinson & Russell, 2007). The testbed uses OAI-PMH to harvest and index data from these repositories to form a unified information space, which can then itself be queried and harvested through standard protocols. To enable this to work smoothly, guidelines were drawn up

10. CLOCKSS Web site, URL: http://www.clockss.org/
11. RepoMMan Project Web site, URL: http://www.hull.ac.uk/esig/repomman/
12. REMAP Project Web site, URL: http://www.hull.ac.uk/remap/
13. DRIVER Project Web site, URL: http://www.driver-repository.eu/

to ensure consistent conventions for metadata content and the use of the OAI-PMH protocol were adopted across the repositories (Vanderfeesten, Summann & Slabbertje, 2008). Services built on top of this information space included a search service and a tool for checking a repository's compliance with the DRIVER guidelines.

Phase Two runs from December 2007 to November 2009, developing the testbed system into a production-quality infrastructure and widening its geographical coverage. Additional services have been added, including profiling and end-user recommendations, and support has been extended to datasets. A DRIVER Network Evolution Toolkit (D-NET) has been developed to make it easier on a technical level to add a repository to the DRIVER information space. It is also tackling the human aspect of implementation with training and community building activities (Lossau & Peters, 2008).

6 Preservation planning tools

According to the OAIS Reference Model (CCSDS, 2002), preservation planning is an activity that encompasses monitoring the repository's environment, reviewing the repository's contents in the light of this monitoring, and drawing up plans for migrating content to new formats or updating the way in which it is rendered or otherwise disseminated to consumers. Three notable tools that assist in preservation planning are presented here: Plato, AONS II and PRONOM-ROAR.

6.1 Plato

Plato is an open-source Web-based tool that supports and automates the process of specifying requirements, evaluating possible solutions and building a plan for preserving a given set of digital objects.[1] Plato implements the Planets preservation planning methodology (Strodl & Becker, 2007) and integrates registries and services for preservation action and characterisation. It is based on earlier work done in the DELOS Digital Preservation Cluster and builds on Utility Analysis to evaluate the performance of various solutions against well-defined requirements and goals. The methodology can be applied to any class of strategy – migration, emulation, normalisation, etc. – and has been validated in a series of case studies.

Plato supports file format identification via DROID, content characterisation via XCL and object comparison via the Planets XCDL service. It supports the use of a template and fragments library for re-using requirements across different sets of objects, and can import and export preservation plans in an XML format (Becker, Kulovits, Rauber & Hofman, 2008).

6.2 AONS II

AONS II (Automated Obsolescence Notification System II) is an open source, platform-independent and configurable tool that automatically provides information from authoritative international registries to support preservation planning.[2] Adapters for querying PRONOM and the Library of Congress Sustainability of Digital Formats Web site are currently available, and more will be written as further registries emerge.

The original aim for AONS was to provide the obsolescence notification service for the PANIC architecture (see section 5.1). With AONS II, the aim was to allow it to

1. Plato Web site, URL: `http://www.ifs.tuwien.ac.at/dp/plato/`
2. AONS II Web site, URL: `http://www.apsr.edu.au/aons2/`

support a national federated infrastructure as well as local standalone repositories and networked enterprise repositories. Specifically, it allows a repository manager to make tailored risk assessments of file formats held in the repository, both at the time of ingest and as a batch operation running on demand or as scheduled. The user can also set up a risk profile for a repository so that when the specified indicators change status, AONS II sends a notification recommending that a risk assessment or preservation action should be performed (Pearson & Walker, 2007).

6.3 PRONOM-ROAR

PRONOM-ROAR is a service that provides file format profiles for over 200 repositories indexed by ROAR, the Registry of Open Access Repositories.[3] ROAR indexes these repositories by harvesting metadata records through OAI-PMH. PRONOM-ROAR generates the file format profiles by downloading all the files associated with those records and identifying their format using DROID. The service has some limitations; for example, it does not include files over 2 MB in size, or files hosted on a different server from the OAI-PMH interface (Brody, Carr, Hey, Brown & Hitchcock, 2007).

3. Preserv Format Profiling: PRONOM-ROAR – an illustrated guide, URL: http://trac.eprints.org/projects/iar/wiki/Profile

7 Metadata

Any system dealing with data has to keep additional information about those data if it is to manage them effectively. Systems that aim to support preservation have very particular metadata needs, needs that have been elaborated by many different projects and initiatives.

One of the most important classes of metadata needed for preservation is what the OAIS Reference Model terms Representation Information, as this is what enables a digital object to be rendered and understood. It is by no means all that is needed, however, so several initiatives have attempted to provide more comprehensive guidance on the metadata to collect. PREMIS concentrates on preservation but aims to be applicable to all digital objects. Other projects and initiatives such as CAIRO focus on particular types of digital object but consider a wider range of applications for metadata. One notable development in this area is the introduction of Dublin Core Application Profiles, which may be scoped according to both content type and application.

Tools for working with such metadata are discussed in chapter 8.

7.1 Representation Information

The OAIS Reference Model contains the concept of a Representation Network, which it defines as: 'The set of Representation Information that fully describes the meaning of a Data Object. Representation Information in digital forms needs additional Representation Information so its digital forms can be understood over the Long Term' (CCSDS, 2002, p. 1-13). This network comes about as pieces of Representation Information are themselves typically Information Objects, consisting of a Data Object and further Representation Information (see Figure 7.1). Recursion is typically halted by a Data Object which can be interpreted by the Designated Community (or in the general case, a particular user) using its Knowledge Base: knowledge, experience and readily available tools.

This recursion, coupled with the fact that the same Information Object can act as Representation Information for multiple other Information Objects, means that bundling a complete set of Representation Information with a Data Object is considerably less practical than giving persistent addresses to pieces of Representation Information and citing them from metadata associated with the Data Object. This is the reasoning behind Representation Information registries – collections of citeable Representation Information that can be used in Representation Networks.

The most mature Representation Information registry is PRONOM, set up by The (UK) National Archives in March 2002.[1] PRONOM holds information about file formats such

1. PRONOM Web site, URL: `http://www.nationalarchives.gov.uk/pronom/`

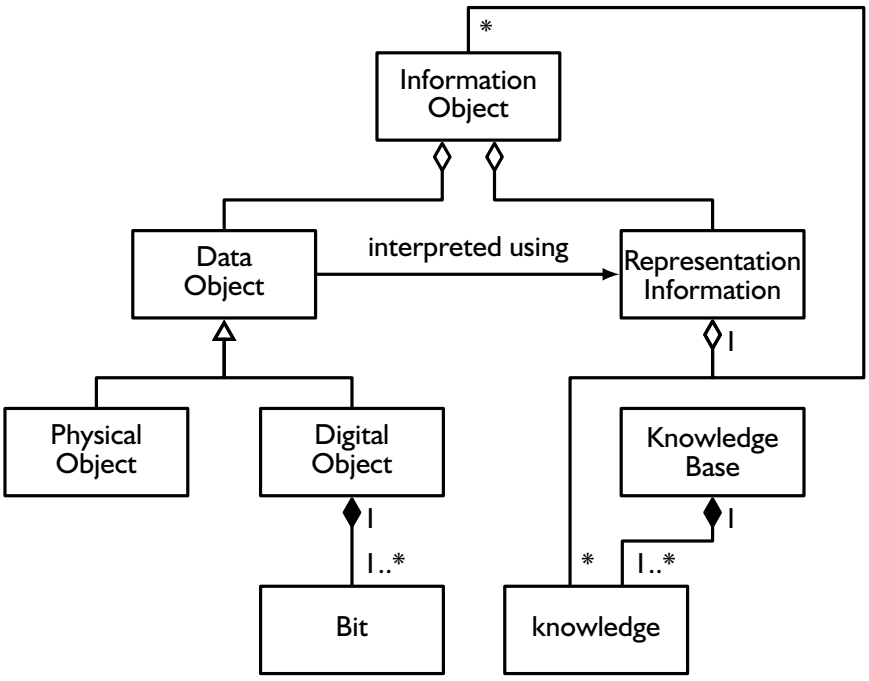

Figure 7.1: A UML class diagram showing Representation Information recursion. All terms except 'knowledge' are as defined in the OAIS Reference Model (CCSDS, 2002, p. 4·20).

as

- their relationships to other formats
- whether they are binary or text-based
- whether they use big-endian or little-endian byte order
- which character encoding they use, if any
- which organisation is responsible for them
- when they were first released, and when support will end (if this has been announced)
- if they have one, the formats' signature (e.g. a bit sequence that all files of this format begin or end with)
- whether there are any rights issues associated with the formats, and
- examples of how they are used in practice.

Each format is given a persistent unique identifier of the form `fmt/`⟨*integer*⟩ which may be appended to the URL `http://www.nationalarchives.gov.uk/pronom/` to retrieve a human-readable presentation of the information held. This information may also be accessed in a machine-readable fashion via the DROID tool.

A more recent initiative is the Global Digital Format Registry (GDFR), of which the first version of the underlying software was delivered in August 2008 after two years of development. This work was accomplished as a collaborative project involving Harvard University Libraries, NARA and OCLC, with funding from the Andrew W.

Mellon Foundation.[2] The aim of the GDFR was to provide a distributed global hierarchy of registries containing both vetted and non-vetted format-related Representation Information (Abrams & Flecker, 2005). The intention was that information vetted by the GDFR review process – including both a public and private review – would be offered as a high quality and authoritative resource, whereas non-vetted information would be provided to give the registry wider coverage. Typical motivations for including non-vetted information include providing early access to information still in the vetting process, allowing local and highly specific information to be registered, and providing incomplete information where little is known about a format.

While it was always intended for PRONOM to become a part of the GDFR, it was felt that the continued development of the two registries in parallel was duplicating effort unnecessarily. Therefore, in April 2009 those involved began a more co-ordinated development approach under a single initiative called the Unified Digital Format Registry (UDFR).[3] The UDFR will have the same broad aims as the GDFR, with a similar approach to quality control, but this time instead of developing a distributed registry system and reconciling it with PRONOM, the initiative will extend the existing PRONOM system and database, adding the necessary features for operating in a distributed manner (UDFR Initiative, 2009).

Another major registry initiative is the Registry/Repository of Representation Information (RRoRI), developed jointly by the Digital Curation Centre and the CASPAR Project (see section 5.4).[4] Unlike PRONOM and the GDFR, RRoRI does not restrict itself to file format information, but also caters for instrument calibrations, data units and other semantic representation information. It therefore uses a data model that concentrates on classifying the representation information, with the information itself held in documents (or software) stored within the repository section of RRoRI (Giaretta, Patel, Rusbridge, Rankin & McIllwrath, 2005).

It is unclear at present how many repositories are making use of these registries beyond those helping to develop them, although the number of projects covered elsewhere in this report that make use of PRONOM via DROID, for example, would suggest they will become increasingly important in future.

7.2 InSPECT

Representation Information is a highly useful concept to use when considering the eventual use of a digital object by a consumer. When a repository is using Representation Information in its own preservation actions, for example when migrating a file from one format to another, it typically has to make a selection from the available Representation Information, and the choice made has an effect on what data will remain available to the consumer. In such circumstances, it is more useful to consider the related concept of a digital object's *significant properties* – those aspects of the digital object itself

2. GDFR Web site, URL: `http://www.gdfr.info/`. GDFR source node, URL: `http://www.formatregistry.org/registry`
3. UDFR Web site, URL: `http://www.udfr.org/`
4. RRoRI Web site, URL: `http://registry.dcc.ac.uk/`

which must be preserved over time in order for the digital object to remain accessible and meaningful. Developing and expounding this concept was the aim of InSPECT (Investigating the Significant Properties of Electronic Content over Time), an 18-month project concluding in March 2009 and led by the Centre for e-Research, King's College London.[5] Four case studies were conducted, in which the significant properties of raster images, emails, structured text and digital audio respectively were defined, and the extent to which these properties survived transformation from one representation format to another were measured. From these case studies, a generic framework for defining significant properties was produced (Knight, 2008b), alongside a significant properties data dictionary and a document discussing the factors that influence decisions on which properties are significant (Knight, 2008a, 2008c).

7.3 PREMIS

The PREMIS Data Dictionary grew out of work by OCLC and RLG to flesh out the preservation metadata required by the OAIS Reference Model with elements from existing schemata (OCLC/RLG Working Group on Preservation Metadata, 2002; Preservation Metadata: Implementation Strategies (PREMIS) Working Group, 2005–05; PREMIS Editorial Committee, 2008). It is uses a data model with five entities: Intellectual Entities, Objects, Events, Agents and Rights. Of these, most of the metadata elements ('semantic units') are associated with the Object entity; none are associated with Intellectual Entities. While PREMIS does not prescribe any particular implementation, XML Schemata are provided for exchanging PREMIS metadata.[6]

Several papers have been written on applying PREMIS in a repository context. Hitchcock, Brody, Hey and Carr (2007b), in the context of the PRESERV service provider model (see section 5.7), map PREMIS metadata elements onto their principal sources, out of the person submitting the content, the repository software, a file format identification/characterisation tool (e.g. DROID), repository policy documents, the preservation service provider, and environment registries. Woodyard-Robinson (2007) describes how PREMIS has been implemented across sixteen different repositories and projects, including the Koninklijke Bibliotheek, the (UK) National Archives, the National Digital Newspaper Program at the Library of Congress, and the National Digital Heritage Archive at the National Library of New Zealand. Dappert and Enders (2008) describe how the British Library used PREMIS alongside METS and MODS as part of the ingest process for electronic journals.

7.4 CAIRO

The CAIRO Project (Complex Archive Ingest for Repository Objects) was a two-year collaboration between the Bodleian Library (University of Oxford), the John Rylands

5. InSPECT Project page at KCL, URL: `http://www.kcl.ac.uk/iss/cerch/projects/completed/inspect`. InSPECT Project Web site, URL: `http://www.significantproperties.org.uk/`
6. PREMIS Web site, URL: `http://www.loc.gov/standards/premis/`

Library (University of Manchester) and the Wellcome Library, ending in August 2008.[7] The aim of the Project was to simplify the task of adding collections of diverse born-digital content to a preservation repository, by integrating a suite of relevant digital curation tools into a simple, unified user interface (Thomas, 2008). As part of the development work, the Project developed an extensible framework of content models for frequently encountered object types, expressed in terms of appropriate descriptive, preservation, structural and administrative metadata to be included in METS files. These content models defined the desired output from the proof-of-concept tool developed by the Project, although not all of them were supported by the conclusion of the Project. The tool was implemented on top of the Eclipse platform,[8] and used a modular architecture to allow tools such as DROID and JHOVE to be plugged in as needed.[9] The tool continues to be developed as part of the futureArch Project.[10]

7.5 Dublin Core Application Profiles

Application profiles are '[metadata] schemas which consist of data elements drawn from one or more namespaces, combined together by implementers, and optimised for a particular local application' (Heery & Patel, 2000, ¶ 2). They are a way of creating a new metadata schema that is explicitly compatible (at least partially) with existing schemata, while also being specifically tailored for a given application. Typically this will involve

- selecting a pertinent subset of elements from an existing schema;
- combining several existing schemata that are each at least partially relevant for the application;
- refining elements from existing schemata, either to provide a narrower meaning in the context of the application, or to control the vocabulary used;
- filling in any gaps with a custom schema, to be separately maintained.

In 2006, the JISC commissioned an application profile to support the Intute Repository Search Project, among the aims of which was to provide a cross-search facility of the scholarly works held in UK institutional repositories (Allinson, Johnston & Powell, 2007). The most mature technology for accomplishing this at the time was OAI-PMH (Open Archives Initiative, 2008), which requires that items are described by a record using the 15 elements of the Dublin Core Metadata Element Set (Dublin Core Metadata Initiative [DCMI], 2004) (though alternative records may also be provided); it made sense, therefore, for the Scholarly Works Application Profile (SWAP) to be based strongly on Dublin Core metadata. The other main influence on SWAP was IFLA's *Functional Requirements for Bibliographic Records* (FRBR), which identifies different levels of abstraction to which metadata attributes may be attached (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998). The finished application profile therefore ended up as five schemata in one, applying to scholarly works, expressions, manifestations, copies

7. CAIRO Project Web site, URL: `http://cairo.paradigm.ac.uk/`
8. Eclipse Web site, URL: `http://www.eclipse.org/`
9. CAIRO ingest tool development page, URL: `http://sourceforge.net/projects/cairo-ingest/`
10. futureArch Project blog, URL: `http://futurearchives.blogspot.com/`

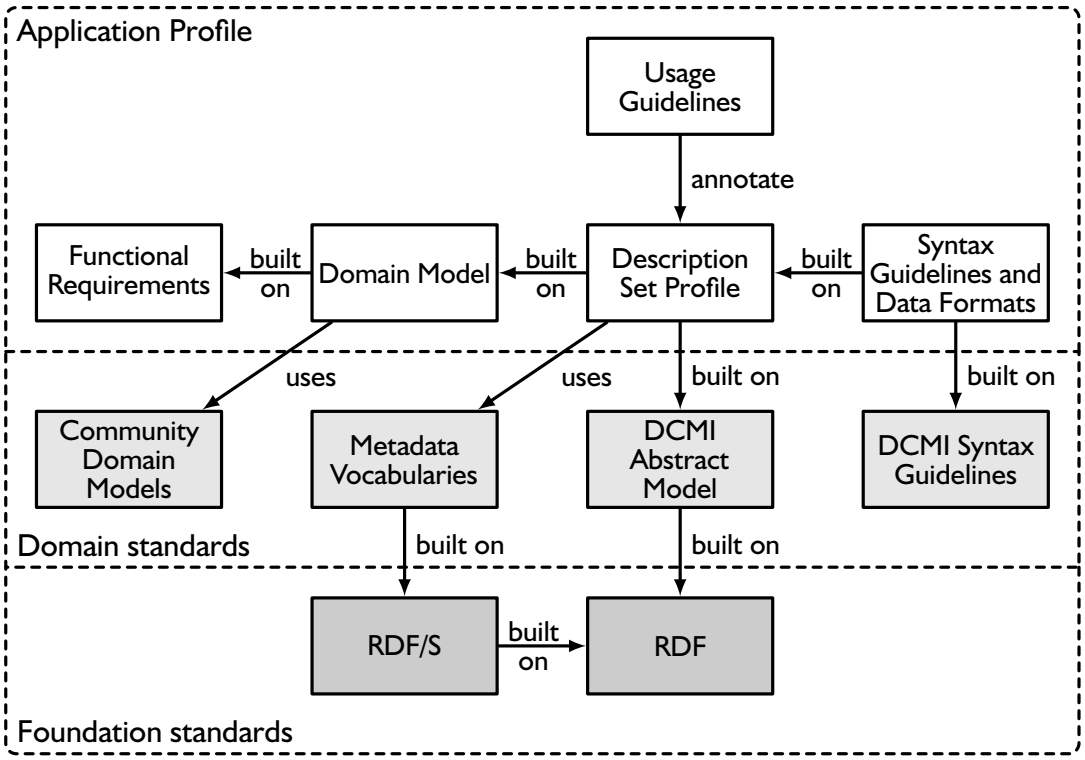

Figure 7.2: Singapore Framework for Dublin Core Application Profiles

and agents respectively. As well as using the Dublin Core Metadata Element Set and the more recent Dublin Core Metadata Terms (Dublin Core Metadata Initiative Usage Board [DCMI Usage Board], 2006), SWAP also contained elements from the Library of Congress schema of MARC Relator terms (Dublin Core Metadata Initiative Usage Board, 2005), and any remaining gaps were filled with a custom schema.[11]

In some ways, SWAP may be seen as a pilot project for using the 'DCMI Abstract Model' (Powell, Nilsson, Naeve & Johnston, 2005) to develop an application profile, resulting in the 'Singapore Framework for Dublin Core Application Profiles' (Nilsson, Baker & Johnston, 2008). This document presents a definition of a Dublin Core Application Profile (DCAP) that is rather more complex and prescriptive than that of application profiles generally (see Figure 7.2). The Framework defines a DCAP as a packet of three to five components:

- *functional requirements (mandatory):* the application that the profile is intended to support, in terms of specific functions that are in (or out) of scope;
- *domain model (mandatory):* the basic entities described by the profile, and the relationships between them.
- *description set profile (mandatory):* a set of rules that define what constitutes a valid instance of the application profile.
- *usage guidelines (optional):* human-orientated information and guidance on using the application profile in practice.

11. Eprints Terms Web page and namespace, URL: `http://purl.org/eprint/terms/`

- *encoding syntax guidelines (optional):* any additional syntactical rules relating to the application that aren't covered by the description set profile and the normal rules of the language used to express instances of the application profile.

In addition, the Framework recommends that the domain model for a DCAP is based on an existing, widely used domain model, and that any syntax guidelines should take into account those provided by DCMI. The Framework requires that the description set profile should specify how entities of the DCMI Abstract Model are used in the DCAP, and by extension how the DCAP may be expressed in RDF. Any metadata vocabularies used by the description set profile should be expressed using RDF Schema, otherwise known as RDF Vocabulary Description Language.

Following on from this, the JISC commissioned several more DCAPs, reflecting the fact that institutional repositories collect more than just scholarly works. Two of the DCAPs were for non-textual materials – still images (Eadie, 2008) and time-based media (audio and video) (Calverley, 2009) – while a third was for the geospatial aspects of any repository object (Reid, 2008). In addition, scoping studies were commissioned to investigate the feasibility and likely benefits and costs of producing DCAPs for learning materials and scientific data (Ball, 2009; Barker, 2008).

In parallel with this work, the JISC has also been developing a registry to enable repositories and other initiatives to register and share their application profiles. The JISC Information Environment Metadata Schema Registry (IEMSR) project began in January 2004, and entered a fourth phase in September 2009.[12] The core partners in the current phase are UKOLN, University of Bath and the Institute for Learning and Research Technology (ILRT), University of Bristol, with contributions from JISC CETIS and the British Educational Communications and Technology Agency (Becta). The eventual aim of the project is to establish IEMSR as a shared service within the JISC Information Environment. Currently, the effort is concentrated on developing three software tools for interacting with the registry data server: a SPARQL (SPARQL Protocol And RDF Query Language) endpoint for machine-to-machine interactions with the registry, a Java-based Web front-end for manually browsing the registry, and a desktop client for building new application profiles, metadata vocabularies, usage guidelines and so on (Tonkin & Strelnikov, 2009).

It should be noted that JISC is not the only body interested in DCAPs. For example, The University of North Carolina Metadata Research Center and the (US) National Evolutionary Synthesis Center have developed a DCAP for use in a repository of scientific data relating to evolution and ecology (Greenberg, White, Carrier & Scherle, 2009).

7.6 Metadata for complex objects

7.6.1 Multipart objects

Open Archives Initiative Object Reuse and Exchange (OAI-ORE) is a standard for identifying and describing aggregations of Web-accessible resources (Open Archives

12. IEMSR Web site, URL: `http://www.ukoln.ac.uk/projects/iemsr/`

Initiative [OAI], 2008). It uses an RDF Resource Map to describe the properties of an aggregation, and identify its constituent parts. It does not describe how the parts relate to one another, although the Resource Map may be extended to include this information; this is because the range of possible relationships is dependent on the types and formats of the resources, and therefore best provided by specialist vocabularies. To cater for situations where the relationships of the constituent parts only apply in the context of a given aggregation, OAI-ORE provides a proxy mechanism for defining a context-laden RDF node for a resource.

7.6.2 Software

Software Preservation: Standards is a research project of the STFC e-Science Centre.[13] It investigates the costs and benefits for software repositories of using standards such as the OAIS Reference Model and PREMIS, while monitoring and helping to develop new and existing standards (Matthews et al., 2009). The project has been informed by a case study looking at the Verified Software Repository (Bicarregui, Hoare & Woodcock, 2006), and how the code for the Repository is itself managed on the SourceForge software repository.[14]

13. *Software Preservation: Standards* Web site, URL: `http://www.e-science.stfc.ac.uk/projects/software-preservation/software-preservation-standard.html`
14. Verified Software Repository development page, URL: `http://vsr.sourceforge.net/`

8 Tools for working with metadata

Several tools have been developed that assist repositories in gathering the metadata they need to support the curation and preservation of the digital objects they contain. JHOVE, DROID and the National Library of New Zealand's Metadata Extraction Tool are tools that enable metadata to be extracted automatically from the digital objects themselves, whereas tools such as SWORD and PREMINT provide friendly ways for repositories to request metadata from those depositing a digital object.

8.1 JHOVE

JHOVE (pronounced 'Jove'), the JSTOR/Harvard Object Validation Environment, is an extensible software framework for performing format identification, validation, and characterisation for digital objects.[1] Determining an object's format, checking the extent to which a file conforms to its format specification, and extracting important metadata from digital objects are frequently necessary during routine operation of digital repositories and for digital preservation activities. These actions are performed by format modules, while the precise form of the reports generated by JHOVE is controlled by output handlers. JHOVE uses an extensible plug-in architecture; it can be configured at the time of its invocation to include whatever specific format modules and output handlers that are desired. Version 1 of JHOVE includes format modules for: arbitrary byte streams; ASCII and UTF-8 encoded text; HTML and XML; GIF, JPEG2000, JPEG and TIFF images; AIFF and WAVE audio; and PDF. Output handlers are provided for text and XML output.

Version 2 of JHOVE is under development at the time of writing.[2] The new version promises to be easier to plug into larger systems, and easier to extend with third party modules. It will no longer use the same process for the identification and validation of formats, it will have a new, configurable assessment action, and will support complex objects (i.e. where a single object may be represented by several files, and files may aggregate streams in different formats). With only limited resources, the JHOVE2 development team has not been able to rewrite all the format modules from version 1, so HTML support will be dropped in favour of the more general SGML, while AIFF, GIF and JPEG will only receive limited attention. Support will be added for ICC colour profiles and shapefiles (Abrams, Morrissey & Cramer, 2009).

1. JHOVE Web site, URL: `http://hul.harvard.edu/jhove/`
2. JHOVE 2 Project wiki, URL: `https://confluence.ucop.edu/display/JHOVE2Info/`

8.2 DROID

DROID (Digital Record Object Identification) is a software tool developed by The National Archives to perform automated batch identification of file formats.[3] It is an open source, platform-independent Java application that may be run in either graphical or command-line mode, and may be plugged into larger systems. It uses the external and internal signatures recorded in the PRONOM registry to determine the formats of individual files; these may be located on the local filesystem, on the Web or streamed directly to the command-line interface. DROID outputs its results to XML, CSV or to the GUI from which it may be printed.

8.3 Metadata Extraction Tool

The Metadata Extraction Tool was developed by the National Library of New Zealand to extract preservation metadata from a range of file formats.[4] The metadata extracted correspond to the elements associated with files in 'Metadata Standards Framework' (National Library of New Zealand, 2003). The tool has a modular architecture, using modules for different file formats; modules are currently provided for: arbitrary byte streams, BMP, GIF, JPEG and TIFF images; MS Word, Excel, PowerPoint and Works documents; WordPerfect, OpenOffice.org and PDF documents; WAV and MP3 audio files; and HTML and XML. The tool may be run in either graphical (Windows) or command-line (UNIX) mode.

8.4 SWORD

SWORD (Simple Web service Offering Repository Deposit) is a lightweight protocol for depositing content in repositories, based on the Atom Publishing Protocol (Gregorio & de hOra, 2007) and making use of the Atom Syndication Format (Nottingham & Sayre, 2005).[5] The SWORD specification was first released on 12 October 2007, and has since gone through several revisions, with version 1.3 released as part of phase 2 of the SWORD project in October 2008 (Allinson et al., 2008). As SWORD is aimed at depositing content, it only modifies Atom's use of HTTP POST requests to create resources; it permits but does not modify Atom's use of PUT and DELETE requests to modify or remove resources, respectively, but instead specifies how SWORD-compliant repositories should react to such requests if that functionality is not supported.

SWORD has requirements for the metadata supplied about a deposit request, but does not specify the metadata that should be supplied about a deposited document. Instead, it requires the deposit client to package the document in a format acceptable to the repository; it is therefore the packaging standard that imposes the document

3. DROID development wiki, URL: http://sourceforge.net/apps/mediawiki/droid/
4. Metadata Extraction Tool Web site, URL: http://meta-extractor.sourceforge.net/
5. SWORD Web site, URL: http://www.swordapp.org/

metadata requirements. A separate specification, SWORD Content Package Types (Downing, 2008), enumerates the packaging standards that deposit clients should support; compliant repositories should therefore accept packages of at least one of these types.

To demonstrate the efficacy of SWORD, demonstrators were implemented both on the client side in the form of deposit clients for the Web, the desktop (in the form of a Java application) and Facebook,[6] and on the repository side with test installations of DSpace, EPrints, Fedora and IntraLibrary.[7] The code for adding SWORD support to existing DSpace, EPrints and Fedora repositories has been made available,[8] as well as a toolkit for building further SWORD deposit Web interfaces. Intrallect has incorporated SWORD support into its IntraLibrary product, while support for depositing documents direct from Microsoft Office applications is available using the OfficeSWORD plugin or Miscrosoft's Article Authoring add-in for Word 2007 (Allinson, François & Lewis, 2008; Currier, 2009).

8.5 PREMINT

PREMINT (PREservation Metadata INput Tool) is an application that gathers preservation metadata from depositors or intermediaries using a graphical interface, and outputs it as a METS XML package.[9] It is tailored for digital artworks, extending previous work by the Guggenheim Museum. Its six components collect, respectively, metadata concerning: Dublin Core elements (title, creator, date created, etc.), technical details of an object's multimedia components, the artistic intention behind the object, information on how the object should be exhibited, the artist's consent for the repository to perform preservation actions, and the structure of the object (expressed in Synchronised Multimedia Integration Language). PREMINT was developed in the context of the PANIC preservation environment (see section 5.1).

6. Web client for deposit via SWORD, URL: `http://client.swordapp.org/`; desktop client for deposit via SWORD, URL: `http://sourceforge.net/projects/sword-app/files/SWORD%20Client/`; client for deposit via SWORD from Facebook, URL: `http://fb.swordapp.org/`
7. DSpace demonstrator, URL: `http://dspace.swordapp.org/`; EPrints demonstrator, URL: `http://sword.eprints.org/`; Fedora demonstrator, URL: `http://fedora.swordapp.org/`; IntraLibrary demonstrator, URL: `http://sword.intralibrary.com/`
8. SWORD downloads page on Sourceforge, URL: `http://sourceforge.net/projects/sword-app/files/`
9. PREMINT Web page, URL: `http://www.itee.uq.edu.au/~eresearch/projects/panic/results/premint.html`

9 Preservation techniques

There are fundamentally three different ways in which digital objects can be preserved, or rather, in which the performances that arise from processes running on the digital objects can be kept reproducible (Heslop, Davis & Wilson, 2002). One is to keep the digital object unchanged, and ensure that the software (or hardware) that interprets it continues to run as intended; this may be accomplished with pure *technology preservation*, or a mixture of this and *emulation*. The second option is to keep the digital object unchanged, and engineer new software (or hardware) to interpret it whenever the existing interpreter becomes unusable; this works best for simple formats, and with low ambitions for the functionality of the interpreter. The third is to accept changes to interpreting software (or hardware), and change the digital object into something a different interpreter can understand; this is *migration*.

A number of tools have been developed that could potentially aid a repository in preserving its contents by these three means.

9.1 Emulation

A typical digital object in a repository is interpreted by a software application, running on an operating system that controls a set of hardware. Emulation typically works by inserting an extra abstraction layer somewhere in this stack, so that the principal concern is ensuring that the abstraction layer runs on as many platforms as possible, rather than the fine detail of how any one particular application works.

Putting the abstraction layer directly between the application layer and operating system layer is possible (e.g. Cygwin translates instructions intended for POSIX systems into ones Windows can understand, while Wine translates instructions intended for Windows into ones POSIX systems can understand) but due to the complexities of modern operating systems, it is not feasible to achieve a perfect mapping. Putting the abstraction layer directly between the operating system layer and the hardware level is also possible in theory, but would likely involve more effort than simply writing new device drivers for the operating system.

Generally the best results come from duplicating one or more layers of the stack; for example, by writing an application that can emulate a hardware platform. This is only an advantage, of course, if the application is written in such a way that it can run on many different platforms.

An example of this is Dioscuri, an emulator developed as part of the Planets Project by the (Dutch) Koninklijke Bibliotheek and Nationaal Archief.[1] It is written in Java, for which

1. Dioscuri Web site, URL: `http://dioscuri.sourceforge.net/`

a runtime environment (or 'virtual machine') is available for most operating systems and architectures. It has three principal parts: a controller, forming the core of the emulator; a library of modules, each of which emulates a particular piece of hardware; and a set of emulator specification documents, each of which lists the modules needed to emulate a particular hardware platform. Development has so far concentrated on emulating a 16-bit x86 PC (Verdegem & van der Hoeven, 2006).

9.2 Reverse engineering

IBM is investigating the most efficient way of using reverse engineering as a preservation approach. For this purpose, IBM is developing its own virtual machine: the Universal Virtual Computer (UVC).[2] While the UVC could eventually be used as a platform for emulator applications such as Dioscuri, it is currently being used in an approach in which custom applications for decoding file formats run directly on the UVC. In this approach, a machine-readable specification called a Logical Data Scheme (LDS) must be written for the digital object's format. The LDS describes in a technology-independent way how the format encodes information, and defines methods for accessing that information. At the time the LDS is written, a software module for the UVC must be written that can decode the format and output information in the way defined by the LDS. Another way of looking at this is that the UVC module is a format decoder, and the LDS documents both the file format and the API (Application Programming Interface) for the decoder. The idea is that the digital object, the UVC module and the LDS are preserved along with the specifications for the UVC for which the module was written. Then, in future, the specifications can be used to rebuild the UVC for whatever platform is current, while the LDS can be used to construct a restore program that uses the UVC module to read the digital object then either renders the object or writes the information to a new file format (Lorie, 2002).

9.3 Migration

9.3.1 Typed Object Model

The Typed Object Model (TOM) is an infrastructure for modelling data types, formats and protocols, thereby allowing conversion between formats or managing similar services through a common interface. It was developed from a proposal by Ockerbloom (1998) at Carnegie Mellon University and the University of Pennsylvania Library. A demonstrator Web service based on TOM and offering conversion between different document formats was made available from University of Pennsylvania Library, though this has since been withdrawn.[3] Versions of the TOM Toolkit for Perl, the TOM Client

2. UVC Web page, URL: http://www.ibm.com/services/nl/dias/cs/uvc.html
3. TOM Conversion Service Web page (archived), URL: http://web.archive.org/web/*/tom.library.upenn.edu/convert/

Toolkit for Java and the TOM Conversion Service software itself are available from the Internet Archive.[4]

There are several Web-based document conversion services still live on the Web, many of which offer a limited free option in addition to a full commercial option.

9.3.2 DExT

The Data Exchange Tools and Conversion Utilities (DExT) Project explored the feasibility of developing tools to aid in the exchange and conversion of primary research data, both quantitative and qualitative.[5] It was led by the UK Data Archive at the University of Essex, and ran from December 2006 to March 2008 (Corti, 2008). The Project had two major outputs: the QuDEx XML schema and the DDI-DExT Tool.

The QuDEx (Qualitative Data Exchange model) XML schema is an attempt to provide a neutral interchange and archival format for Computer Assisted Qualitative Data Analysis Software (CAQDAS) packages. The model was developed in consultation with the social science data archiving community, XML schema experts and CAQDAS vendors. It packages together code, classification, memo and document data, along with a few other types of resource, but does not provide for the full gamut of materials and metadata; it is suggested that a standard such as METS is used to package this latter information with the QuDEx data. Version 3.0 of the model and schema were released in February 2008. Following the conclusion of DExT, QuDEx is now maintained by the Data Documentation Initiative (DDI) committee and supported by the Open Data Foundation.

The DDI-DExT Tool was developed in collaboration with the Open Data Foundation between June 2007 and February 2008. It provides a means to convert statistical data between commonly used formats and a neutral archival format. DDI-DExT supports SPSS as an input format, and various versions and flavours of SAS, Stata and SPSS as output formats. The neutral archival format uses fixed ASCII for data and DDI version 3.0 for metadata. DDI-DExT is open source, and was implemented on top of the Eclipse platform.[6] The initial release was intended as a proof of concept, and thus is not quite ready for production use (Heus, 2008).

4. TOM Software Web page (archived), URL: `http://web.archive.org/web/*/tom.library.upenn.edu/sw/`
5. DExT Web site, URL: `http://www.data-archive.ac.uk/dext/`
6. DDI-DExT development page, URL: `http://opendatafoundation.org/?lvl1=forge`

10 Identifiers

Many of the technologies used for curating and preserving digital objects rely on objects being given unique identifiers. For the most part these only need to be locally unique, but as document and data repositories join up in various ways there is an advantage to objects having globally unique identifiers. There is a tendency, therefore, among global identification schemes to define identifiers in two parts: the first part identifying an agency, and the second part being that agency's local identifier for an object.

In terms of mere identification, all that is required of a scheme is that it is generates identifiers that are unique in a given context. So long as the identifier remains unique, it remains useful as an identifier while the scheme is known, and is in that sense persistent. There is natural tendency, though, to burden identifiers with tasks other than mere identification. Such tasks may include conveying information about the identified object, such as a hint as to its title or subject matter, or conveying by comparison the relationships between identified objects, for example that one object is part of a larger object, or that two objects are different versions of one another. Perhaps the most important of these secondary purposes is to provide a locator for the object, so that a user can enter the identifier into a retrieval system such as a Web browser and bring up a copy of it. It is this last point that elicits most concern about identifier schemes, as keeping locators persistent implies preserving a great deal more infrastructure than merely keeping identifiers unique (Bellini, Cirinnà & Lunghi, 2008; Hilse & Kothe, 2006).

The point about infrastructure is key, at least when it comes to global identifiers. If a global registrar requires consultation as each identifier is minted, this creates a single point of failure for the entire scheme, whereas if minting identifiers is delegated to individual agencies, the agencies can continue to do so even if the global registrar suffers difficulties. The cost of minting an object identifier or maintaining an agency identifier will also have an impact on the sustainability of the identifier infrastructure and its usage. Furthermore, identification infrastructures with well-defined business models, exit strategies and commitments to uniqueness are more likely to inspire trust than those without.

10.1 Common identifier schemes

The scheme by which resources are identified on the Internet is the Uniform Resource Identifier (URI) scheme. The URI scheme is actually an umbrella term for several different identification schemes, all of which begin with a scheme identifier or protocol name marked off by a colon. The official register of such schemes is maintained by the Internet Corporation for Assigned Names and Numbers (ICANN), although there are many unregistered schemes in existence (Mealling & Denenberg, 2002'Uniform Resource Identifer (URI) Schemes per RFC4395', 2010).

A subset of URIs have an associated protocol for retrieving the identified resources; these schemes are known as Uniform Resource Locators (URLS). The most well known of these are HTTP URLs (`http:`) though there are several others: `https:`, `ftp:`, `file:`, `gopher:` and so on. HTTP URLs are frequently criticised for not being persistent, either from a locator point of view or from an identifier point of view. This is partly because the scheme has no commitment to persistence at a global level – domain names and IP addresses that are no longer in use may be reused by another host entirely. Neither is there a culture at the host level of keeping the paths persistent and unique, despite encouragement and occasional exemplary practice to the contrary (Sauermann & Cyganiak, 2008). Nevertheless, the convenience that HTTP URLs afford for retrieving documents means that many other identifier schemes have an HTTP URL expression, even if their canonical expression is not an HTTP URL.

There are several services that provide URL redirection on the Web, either expressly to protect links from breaking when the real URL of a resource changes, or to shorten an otherwise unwieldy URL. Of the former, the most well-known is the PURL (Persistent URL), established by OCLC and in continuous operation since 1996.[1] The PURL resolver software is freely available, allowing other organisations to set up their own resolver.[2]

The Handle System was developed by the Corporation for National Research Initiatives with funding from the (US) Defence Advanced Research Projects Agency, and has been running since 1994.[3] The canonical form of the Handle identifier is a code representing a naming authority (usually a pair of numbers separated by a dot) followed by a slash, followed by a locally assigned identifier; the official way of representing this as a URI is to prefix it by '`hdl:`'. The system operates its own system of resolvers, meaning it is possible to retrieve information about an identified resource independent of any other Internet application, but an HTTP proxy server is provided which will redirect URLs of the form `http://hdl.handle.net/`⟨*handle*⟩ to a URL in the Handle record for the resource. The proxy supports load balancing between several URLs.

The root naming authority `10` under the Handle System has been assigned to the International DOI Foundation (IDF), with all Handles assigned by it being known as Digital Object Identifiers (DOIs). The DOI system is aimed at journal publishers; those participating as registration agencies are given a naming authority code of the form `10.`⟨*number*⟩, and are responsible for assigning unique codes for each journal article (say) they publish and storing appropriate metadata for the identified object within the Handle system. DOIs may use an alternative URI representation (beginning '`doi:`') and a dedicated HTTP proxy server is provided for use with DOIs, as an alternative to the main Handle proxy.[4]

The Archival Resource Key (ARK) system was developed by Kunze and Rodgers (2008) and is currently maintained by the California Digital Library. The canonical form of the identifier is a URI beginning with `ark:/`, followed by a five- or nine-digit number representing a Name Assigning Authority, followed by another slash and the object name. The name may be qualified by adding a component path (introduced and subdivided

1. OCLC PURL resolver, URL: `http://purl.oclc.org/`
2. PURL Community Web site, URL: `http://www.purlz.org/`
3. Handle System Web site, URL: `http://handle.net/`
4. DOI HTTP proxy server, URL: `http://dx.doi.org/`

by slashes) indicating finer granularity, and a variant path (introduced and subdivided by dots) indicating a specific version of the resource. While the identifier itself refers to a specific resource, the identifier followed by '?' refers to metadata about that resource, and the identifier followed by '??' to a commitment statement from the current provider about curating the resource. An ARK is (currently) made resolvable by prepending it with `http://` and the hostport of a Name Mapping Authority capable of resolving the identifiers of the Name Assigning Authority; this part is expected to vary over time, and mechanisms are proposed for recovering from a faulty Name Mapping Authority specification.

The National Bibliographic Number (NBN) is a namespace within the Uniform Resource Name (URN) scheme (Hakala, 2001). Child namespaces within NBN for the most part take the form of a two-letter country code (ISO 3166-1, 2006), and are allocated to the respective national library and registered with the Library of Congress. The Deutsche Nationalbibliothek provides a resolution service; NBN URNs may be converted to URLs by prepending `http://nbn-resolving.de/`.

10.2 PILIN

The PILIN Project (Persistent Identifier and Linking Infrastructure), run by Link Affiliates, developed a common infrastructure for the persistent identification of digital assets produced in the Australian higher education community and held in cultural heritage institutions.[5] In the 16 months from September 2006, the project produced best practice guides on the use of persistent identifiers, a set of use cases, an ontology for describing persistent identifier infrastructures, and a pilot infrastructure based on the Handle System. Several demonstrator tools were developed: a simple JAVA API for interacting with the Handle System, a Web interface for managing Handle records, a tool for using the FRBR model (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998) with Handles, and a reverse lookup service for finding the Handle associated with a URL, among others (PILIN Team, 2007).

A second phase of PILIN, known as the PILIN ANDS Transition Project, ran from January to June 2008 to plan for turning the pilot infrastructure of PILIN into an Australian national persistent identifier service under the auspices of the Australian National Data Service (ANDS).[6] It was hoped that a service provider would be selected at the conclusion of the project, but this turned out not to be the case. Since the conclusion of this project, Link Affiliates have continued to refine the PILIN identification ontology, and to work on policies for maintaining identifier persistence.

5. PILIN Project Web site, URL: `http://www.pilin.net.au/`
6. PILIN ANDS Transition Project Web site, URL: `http://linkaffiliates.net.au/pilin2/outputs/outputs_idmodelling.html`

10.3 RIDIR

The RIDIR Project (Resourcing IDentifier Interoperability for Repositories) investigated the requirements for, and benefits of, using persistent identifiers that work across many digital repositories of different types.[7] The Project ran from April 2007 for one year, and was lead by the University of Hull with development effort from Rightscom Ltd. The Project produced several use cases for identifiers – de-duplication of resources, resolution to the version of record regardless of current location, and discovery of related resources in different locations – and a demonstrator to show how the latter two might be achieved in practice (Green, 2008).

7. RIDIR Project Web site, URL: http://www.hull.ac.uk/ridir/

11 Guidelines, Certification and Audit

While having a workable tool set for dealing with the technical side of preserving and curating repository content is vital, a repository must also be managed well if it is to be effective. To this end, tools have also been developed to help with drawing up policies and procedures for repositories, and for checking the robustness of the service provided by a repository.

The Data Audit (or Asset) Framework enables a repository to map out the kinds of data asset it needs to support, while SHERPA/ROMEO, the OpenDOAR Policies Tool and the DRIVER Guidelines for Content Providers provide guidance for drawing up repository policies and procedures. Finally, audit tools such as the DINI-Zertifikat, DRAMBORA, TRAC and PLATTER allow the trustworthiness and robustness of a repository to be assessed, and thereby highlight aspects of the repository and its supporting policy framework that need to be adjusted or improved.

11.1 Data Audit Framework

Inspired by the DRAMBORA methodology for repository assessment (see section 11.6), the Data Audit Framework (DAF) was developed as a means for Higher Education Institutions to recognise their data assets and, as a result, reconsider their policies, procedures and practices for managing them.[1] Being a framework, the DAF does not provide a strict method for performing a data audit, but rather a methodology that can be adapted for different institutional contexts. The DAF involves four stages: *planning the audit* (preliminary research, constructing the detailed method), *identifying and classifying assets* (by means of desk research, interviews and/or questionnaires), *assessing the management of data assets* (gathering metadata, determining current management practice and considering alternatives), and *reporting and recommendations* (presenting findings and indicating what improvements could be made). Sample forms are provided for paper-based audits, and an online tool is provided for completing the audit electronically (Jones, 2009; Jones, Ross, Ruusalepp & Dobreva, 2009). The tool will in future be known as the Data Asset Framework.

1. Data Audit Framework Web site, URL: `http://www.data-audit.eu/`

11.2 RoMEO

The Rights Metadata for Open Archiving (RoMEO) Project, conducted by the University of Loughborough in the 12 months from August 2002, investigated copyright and other rights issues surrounding the self-archiving of research publications, and enumerated the metadata that repositories need to hold in order to fulfil their obligations in this regard (Oppenheim, Probets & Gadd, 2003). This research was followed up by the JISC/SURF Partnering on Copyright Programme,[2] while the database of publishers' self-archiving policies produced by RoMEO was developed into a service for repositories by the SHERPA Project.[3] The service provides for a wide range of journals summary information on which versions of an article an author may self-archive, alongside other pertinent information such as embargo periods and whether the publisher's policy coincides or conflicts with funder policies on open access; where available, a link is supplied to the publisher's policy.

11.3 OpenDOAR Policies Tool

The OpenDOAR Policies Tools was developed in response to research indicating that two thirds of Open Access repositories did not have policies for content submission, re-use, preservation, etc. (Millington, 2006). It provides a set of templates for metadata, data, content, submission and preservation policies, and a simple interface for choosing between the given alternatives.[4] Policies created using the tool can be output as plain text, HTML, or as EPrints source code for static pages or configuration files. The latter files allow the policies to be accessed through the OAI-PMH protocol.

11.4 DRIVER Guidelines for Content Providers

DRIVER (see section 5.11) has produced a set of guidelines for digital repositories participating in the DRIVER infrastructure (Vanderfeesten et al., 2008). The aim of these guidelines is to harmonise the way in which OAI-PMH is implemented across DRIVER repositories, so that the data thus made available is sufficiently uniform to support rich search and retrieval services. The guidelines consist of five main components:

- *Collections.* Resources should be placed in different sets according to whether full text is available from the repository, and whether the resource is openly accessible.
- *Metadata.* Specifications are provided for the use of Dublin Core Metadata Elements in the default OAI-PMH records. DRIVER encourages the use of additional, more comprehensive metadata schemes such as Dublin Core Metadata Terms.

2. Partnering on Copyright Programme Web site, URL: `http://www.lboro.ac.uk/departments/dis/disresearch/poc/`
3. SHERPA/RoMEO Service introductory Web page, URL: `http://www.sherpa.ac.uk/projects/sherparomeo.html`
4. OpenDOAR Policies Tool Web page, URL: `http://www.opendoar.org/tools/en/policies.php`

- *OAI-PMH implementation.* Specifications are provided for representing repository data according to OAI-PMH version 2.0. DRIVER also makes recommendations for handling certain OAI-PMH requests and repository events.
- *Vocabularies and semantics.* Unambiguous vocabularies are provided for document types and versioning.
- *Best practice.* Specifications are provided on various other matters, such as quality labels, persistent identifiers, usage statistics and intellectual property rights.

11.5 DINI-Zertifikat

The Deutsche Initiative für Netzwerkinformation (DINI) Certificate Document and Publication Services (DINI Electronic Publishing Working Group, 2006) provides a standard of quality for higher education institutional repositories. It provides both minimum standards and recommendations for the visibility of services; repository policies; support provided for authors; handling of copyright and licensing issues; security, authenticity and data integrity of both the repository and individual documents; subject indexing, metadata export and repository interfaces; logs and statistics; and long-term availability. The certificate itself is awarded after external inspection.

11.6 DRAMBORA

DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) is a self-assessment toolkit developed by the Digital Curation Centre and DigitalPreservationEurope.[5] The toolkit may be used interactively online or downloaded for use in paper form. The audit has six stages, in which the auditor identifies the organisation's role and objectives, policy framework, activities and assets, before identifying and assessing the risks associated with these activities and assets, and developing a strategy to manage them (Innocenti & Vullo, 2009).

11.7 TRAC

Trusted Repositories Audit and Certification (TRAC) has its origins in a 2002 report on the attributes and responsibilities of trusted digital repositories, written by a joint working group of OCLC and RLG (RLG/OCLC Working Group on Digital Archive Attributes, 2002). This latter report fulfilled its brief at a rather abstract level, and recommended further work to develop a certification programme using detailed and specific criteria. As a result, RLG and the (US) National Archives and Records Administration (NARA) set up an international task force for digital repository certification, producing a draft audit checklist in 2005 (RLG-NARA Digital Repository Certification Task Force, 2005)

5. DRAMBORA Web site, URL: `http://www.repositoryaudit.eu/`

and a final version of the TRAC Criteria and Checklist in 2007 (RLG-NARA Digital Repository Certification Task Force, 2007). The criteria are divided into three sections, relating to organisational infrastructure (governance, sustainability, staffing, etc.), digital object management (OAIS functions), and technical infrastructure (system architecture, technologies, security) respectively. TRAC is now maintained by the Center for Research Libraries (CRL).

11.8 PLATTER

The Planning Tool for Trusted Electronic Repositories (PLATTER) is a strategic-level tool devised by DigitalPreservationEurope for ensuring trustworthiness emerges as a characteristic of a digital repository (Rosenthal et al., 2008). The tool begins with a questionnaire for determining the character of the repository in question, then provides a framework for defining goals and performance targets for the repository, and advice on expanding these into a set of nine Strategic Objective Plans. These plans cover, respectively, financial monitoring, staffing, data/metadata, acquisition of content, access to content, preservation of content, technical systems, succession and disaster planning. The aim for PLATTER is that a repository developing and implementing its own version of the Strategic Objective Plans should find itself performing well in trustworthiness audits.

12 Costs and benefits

For curation, and preservation in particular, to be effectively undertaken requires a not inconsiderable outlay in terms of both staff time and equipment. It is also a long-term commitment on the part of an institution; indeed, it should form part of the institution's core infrastructure (Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010, p. 83). If curation activities are to receive the sustained investment needed, institutions need to be convinced first of all that there is either a net financial benefit to performing curation or a worthwhile return for the additional cost, and second of all that the chosen architecture for curation represents the best value for money for the institution. Making either case requires a framework for assessing the costs and benefits of curation. The four main frameworks developed so far for this purpose are *espida*, LIFE, Keeping Research Data Safe and the *Identifying Benefits* report.

12.1 ESPIDA

The *espida* (Effective Strategic model for the Preservation and disposal of Institutional Digital Assets) Project ran from October 2004 to January 2007, and in that time developed an approach to the process of funding projects.[1] In the higher and further education context in particular, there is a tendency for projects to produce new knowledge, new information or improved processes rather than financial rewards. This makes it hard for proposers to present a project's risks and rewards in anything other than vague terms, and makes it hard for the funders to set these risks and rewards against the requested investment. The *espida* Approach is an attempt to provide a more objective and concrete way of expressing the intangible outcomes of projects. It encourages proposers to align their proposals to the funder's strategic goals, and uses Outcome Scorecards and cost templates to express benefits and costs. The *espida* Handbook includes a case study using the Approach in an institutional repository setting (Currall & McKinney, 2007, pp. 36–42).

12.2 LIFE

The LIFE (Lifecycle Information For E-literature) Project is a collaboration between University College London (UCL) and the British Library.[2] It is looking at the lifecycle of digital material, with particular regard to the cost of collecting and preserving the material.

1. *espida* Project Web site, URL: `http://www.gla.ac.uk/espida/`
2. LIFE Project Web site, URL: `http://www.life.ac.uk/`

The first phase of the Project ran for one year ending in April 2006; it used UCL and British Library collections to develop a model of the lifecycle of a digital object, and a methodology for costing each stage in that lifecycle (McLeod, Wheatley & Ayris, 2006). Within the second phase, which ran for 18 months and ended in August 2008, the lifecycle model was refined and three further costing case studies were produced: SHERPA-LEAP, SHERPA-DP and the Burney Collection (Ayris et al., 2008). The third phase began in August 2009 and will last one year. The focus of this phase will be the further refinement of the Generic Preservation Model and to develop a predictive costing tool, available both as a spreadsheet and as a Web application.

12.3 Keeping Research Data Safe

Beagrie, Chruszcz and Lavoie (2008) propose a framework for determining the medium to long term costs to higher education institutions (HEIs) of preserving research data.[3] The framework uses full economic costing, in order to support more accurate cost–benefit analysis and more accurate comparisons between in-house and outsourced solutions, and is tailored towards use with the Transparent Approach to Costing (TRAC) in use in UK HEIs. It consists of three parts: a list of key cost variables and units that affect how preservation costs change over time; an activity model identifying the activities required to preserve research data, each of which having a cost implication; and a resources template providing TRAC-orientated cost category headings into which the costs identified by the activity model should be separated. The framework was developed with reference to the LIFE cost models (see section 12.2), the OAIS model (see chapter 4) and the NASA Cost Estimation Tool (Fontaine, Hunolt, Booth & Banks, 2007), alongside four case studies of working data archives.

The framework proposed by Beagrie et al. was further refined in the Keeping Research Data Safe 2 Project, which ran for ten months from March 2009. A wider data survey was conducted to corroborate the findings of the earlier report, and the activity model was reviewed and modified accordingly.[4]

12.4 Identifying Benefits

Fry, Lockyer, Oppenheim, Houghton and Rasmussen (2008) identify the benefits arising from curating and openly sharing research data.[5] Among the direct benefits identified were: the potential for new discoveries from existing data, reduced duplication of data collection costs, increased transparency of the scientific record, and higher and more rapid impact of scientific research. Fry et al. also identified potential for pedagogical uses of research data as an indirect benefit.

3. *Keeping Research Data Safe* project page on the JISC site, URL: `http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx`
4. Keeping Research Data Safe 2 Project Web page, URL: `http://www.beagrie.com/jisc.php`
5. *Identifying Benefits* project page on the JISC site, URL: `http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/databenefits.aspx`

Their report contains a methodology for performing a cost–benefit analysis of curating research data. On the costs side it extends the KRDS model (Beagrie et al., 2008) to take account of costs incurred by those depositing and re-using data. On the benefits side, it focuses on calculating direct and indirect cost savings, while providing some scope for calculating increased returns and other more distant benefits. Appendix 3 of the report provides a worked example of a cost–benefit analysis for an institutional data repository.

13 Conclusions

> In a budget crunch, the institutional repository may be one of the last things that can be cut, given the way that digital preservation demands steady and consistent attention and hence funding.
>
> Lynch (2003)

A survey of 21 repositories in 2006 revealed that none of them had a formal preservation policy; there were, however, policies on acceptable file formats and performing format migrations on ingest (Hitchcock et al., 2007c). Since that time, a great deal of research and advocacy has taken place, notably the JISC-funded Digital Preservation Policies Study (Beagrie, Semple, Williams & Wright, 2008) and the DCC-RIN Research Data Management Forum. The situation has improved somewhat, but even so there is much progress still to be made before preservation and curation are fully integrated into the work of institutional repositories: at the time of writing, the Directory of Open Access Repositories reports that 33 (20%) of the 169 UK repositories it monitors specify a preservation policy in their OAI-PMH Identify response.[1]

Happily, the state of the art is moving on apace. On the technical side, tools for performing preservation tasks are maturing. Projects such as CASPAR, DRIVER, Planets and SHAMAN are developing a range of useful tools for the entire curation lifecycle, while SHERPA DP and the CRiB demonstrate how such tools may be integrated to form a comprehensive toolkit. These developments are, crucially, filtering through into popular repository software, with the contributions of PRESERV and KeepIt to EPrints being just one example. On the more strategic and procedural side, audit and certification tools have reached the level where they may be used in earnest, with software helping to ease the use of otherwise paper-based methods. Robust procedures are now available for judging the costs and benefits of preservation activity, allowing the case for preservation and curation to be put in institutions that have not been convinced of the need.

The prospects for preservation and curation in institutional repositories are therefore positive. It is true that the current state of practice is not ideal, and that it will be quite some time before institutional repositories complete the transition from being tools of immediate access to trustworthy, long-term guardians of important digital assets. This transition is visibly underway, though, and provided that the momentum exhibited by the curation and repository communities is not lost, one can be cautiously optimistic about its completion in due course.

1. OpenDOAR Web site: 'Recorded Preservation Policies – United Kingdom', URL: `http://www.opendoar.org/onechart.php?cID=224&potID=5&groupby=pog.pogHeading&orderby=pog.pogID&charttype=pie&width=600&height=300&caption=Recorded%20Preservation%20Policies%20-%20United%20Kingdom`

Bibliography

Abrams, S. L. & Flecker, D. (2005). *A proposal for a Global Digital Format Registry*. Retrieved from http://hul.harvard.edu/gdfr/documents/Proposal-2005-09-29.doc

Abrams, S., Morrissey, S. & Cramer, T. (2009). 'What? So what': The next-generation JHOVE2 architecture for format-aware characterization. *International Journal of Digital Curation*, *4*(3), 123–136. Retrieved February 19, 2010, from http://www.ijdc.net/ijdc/article/view/139

Allinson, J. (2006, November 21). *OAIS as a reference model for repositories: An evaluation*. University of Bath, UKOLN. Retrieved June 25, 2009, from http://www.ukoln.ac.uk/repositories/publications/oais-evaluation-200607/

Allinson, J., Carr, L., Downing, J., Flanders, D. F., François, S., Jones, R., Lewis, S., Morrey, M., Robson, G. & Taylor, N. (Eds.). (2008, October 7). *SWORD AtomPub profile*. Version 1.3. Retrieved February 11, 2010, from the SWORD Project Web site: http://www.swordapp.org/docs/sword-profile-1.3.html

Allinson, J., François, S. & Lewis, S. (2008, January). SWORD: Simple Web-service Offering Repository Deposit. *Ariadne*, *54*. Retrieved February 11, 2010, from http://www.ariadne.ac.uk/issue54/allinson-et-al/

Allinson, J., Johnston, P. & Powell, A. (2007, January). A Dublin Core Application Profile for scholarly works. *Ariadne*, *50*. Retrieved November 23, 2009, from http://www.ariadne.ac.uk/issue50/allinson-et-al/

Asset store. (2009, February 9). Retrieved February 19, 2010, from the DSpace Foundation Web site: http://wiki.dspace.org/index.php?title=AssetStore&oldid=11324

Ayris, P., Davies, R., McLeod, R., Miao, R., Shenton, H. & Wheatley, P. (2008, August 26). *The LIFE*^{2} *final project report*. JISC. Retrieved October 13, 2009, from http://eprints.ucl.ac.uk/11758/

Ball, A. (2009, June 3). *Scientific data application profile scoping study*. JISC. Retrieved November 24, 2009, from http://www.ukoln.ac.uk/projects/sdapss/papers/ball2009sda-v11.pdf

Barker, P. (2008, December 9). *Learning material application profile scoping study*. JISC. Retrieved November 24, 2009, from http://www.icbl.hw.ac.uk/lmap/lmapscopingreport.pdf

Bass, M. J., Stuve, D., Tansle, R., Branschofsky, M., Breton, P., Carmichael, P. et al. (2002, March 1). *DSpace internal reference specification: Technology and architecture*. Retrieved from the DSpace Foundation Web site: http://www.dspace.org/technology/architecture.pdf

Beagrie, N. (2006, November). Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation*, *1*(1). Retrieved November 16, 2009, from http://www.ijdc.net/ijdc/article/view/6

Beagrie, N., Chruszcz, J. & Lavoie, B. (2008, April). *Keeping research data safe: A cost model and guidance for UK universities*. HEFCE. Retrieved October 13, 2009, from http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

Beagrie, N., Semple, N., Williams, P. & Wright, R. (2008, October). *Digital preservation policies study*. JISC. Retrieved February 22, 2010, from http://www.jisc.ac.uk/publications/reports/2008/jiscpolicyfinalreport.aspx

Becker, C., Ferreira, M., Kraxner, M., Rauber, A., Baptista, A. A. & Ramalho, J. C. (2008, September). Distributed preservation services: Integrating planning and actions. In B. Christensen-Dalsgaard, D. Castelli, B. A. Jurik & J. Lippincott (Eds.), *Research and advanced technology for digital libraries: Proceedings of the 12th European Conference on Digital Libraries (ECDL'08)* (Vol. 5173, pp. 25–36). Lecture Notes in Computer Science. Springer-Verlag. Aarhus, Denmark. Retrieved October 26, 2009, from http://hdl.handle.net/1822/8239

Becker, C., Kulovits, H., Rauber, A. & Hofman, H. (2008, June). Plato: A service oriented decision support system for preservation planning. In R. Larsen, A. Paepcke, M. Naaman & J. Borbinha (Eds.), *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries(JCDL'08)* (pp. 367–370). New York, NY: Association for Computing Machinery. Pittsburgh, PA. doi:10.1145/1378889.1378954

Bellini, E., Cirinnà, C. & Lunghi, M. (2008, June). *Persistent identifiers for cultural heritage*. DigitalPreservationEurope. Retrieved February 16, 2010, from http://www.digitalpreservationeurope.eu/publications/briefs/persistent_identifiers.pdf

Bicarregui, J., Hoare, C. A. R. & Woodcock, J. (2006). The verified software repository: A step towards the verifying compiler. *Formal Aspects of Computing*, *18*(2), 143–151. doi:10.1007/s00165-005-0079-4

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010, February). *Sustainable economics for a digital planet: Ensuring long-term access to digital information*. Retrieved March 9, 2010, from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Bodhmage, K. (2005, December 1). *Fedora architectural review*. AHDS. Retrieved February 22, 2010, from http://web.archive.org/web/*/www.sherpadp.org.uk/documents/wp56-fedora.pdf

Bose, R. & Reitsma, F. (2006, May). *Advancing geospatial data curation* (Institute of Geography Online Paper GEO-013). University of Edinburgh, School of Geosciences. Retrieved November 11, 2009, from http://hdl.handle.net/1842/1074

Brody, T., Carr, L. & McSweeney, P. (2010, February 8). *Preservation support*. Retrieved February 19, 2010, from the University of Southampton Web site: http://wiki.eprints.org/w/Preservation_Support

Brody, T., Carr, L., Hey, J. M. N., Brown, A. & Hitchcock, S. (2007, December). PRONOM-ROAR: Adding format profiles to a repository registry to inform preservation services. *International Journal of Digital Curation*, *2*(2). Retrieved November 19, 2009, from http://www.ijdc.net/ijdc/article/view/53/

Brown, A. (2007). Developing practical approaches to active preservation. *International Journal of Digital Curation*, *2*(1). Retrieved February 15, 2010, from http://www.ijdc.net/ijdc/article/view/37/

Calverley, G. (2009, October 29). *Time-based media application profile to support search and discovery*. Retrieved November 24, 2009, from the University of Manchester Web site: http://wiki.manchester.ac.uk/tbmap/index.php/Main_Page

CASPAR Consortium. (2007, May 17). *CASPAR overall component architecture and component model* (Report CASPAR-D1301-TN-0101-1_1). CASPAR Project. Retrieved October 21, 2009, from http://www.casparpreserves.eu/Members/cclrc/Deliverables/caspar-overall-component-architecture-and-component-model-1/at_download/file

CASPAR Consortium. (2009, January 15). *CASPAR draft testbed implementation plan* (Report CASPAR-4105-RP-0101-1_2). CASPAR Project. Retrieved October 21, 2009, from http://www.casparpreserves.eu/Members/metaware/Deliverables/caspar-draft-testbed-implementation-plan/at_download/file

Celeste, E. & Branschofsky, M. (2002). Building DSpace to enhance scholarly communication. In W. Jones (Ed.), *E-serials: Publishers, libraries, users and standards* (2nd ed., pp. 239–247). New York: Haworth Press. Retrieved from http://www.dspace.org/news/articles/celeste.pdf

Consultative Committee for Space Data Systems. (2002). *Reference model for an Open Archival Information System (OAIS)*. Blue Book. Also published as ISO 14721:2003. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf

Corti, L. (2008, March). *Data Exchange Tools and Utilities (DExT)*. JISC. Retrieved February 19, 2010, from http://www.data-archive.ac.uk/dext/reports/DExT_Finalreport_JISC.doc

Currall, J. & McKinney, P. (2007, January 23). *Espida handbook: Expressing project costs and benefits in a systematic way for investment in information and IT*. University of Glasgow. Retrieved October 13, 2009, from http://hdl.handle.net/1905/691

Currier, S. (2009, February 2). *SWORD: Cutting through the red tape to populate learning materials repositories*. JISC CETIS. Retrieved February 11, 2010, from http://www.elearning.ac.uk/features/sword

Dappert, A. & Enders, M. (2008). Using METS, PREMIS and MODS for archiving eJournals. *D-Lib Magazine*, *14*(9/10). doi:10.1045/september2008-dappert

Digital Curation Centre. (2007, April 26). *What is digital curation?* Retrieved November 11, 2009, from http://www.dcc.ac.uk/about/what/

DINI Electronic Publishing Working Group. (2006, September). *DINI-Certificate Document and Publication Services 2007*. DINI Schriften. Version 2.0. Retrieved February 17, 2010, from the Deutsche Initiative für Netzwerkinformation Web site: http://edoc.hu-berlin.de/series/dini-schriften/2006-3-en/PDF/3-en.pdf

Downing, J. (Ed.). (2008, November 6). *SWORD content package types*. Version 1.0 (draft). Retrieved February 11, 2010, from the SWORD Project Web site: http://www.swordapp.org/docs/sword-type-1.0.html

Dublin Core Metadata Initiative Usage Board. (2005, December 8). *MARC Relator terms and Dublin Core*. Dublin Core Metadata Initiative. Retrieved November 23, 2009, from http://dublincore.org/usage/documents/relators/

Dublin Core Metadata Initiative Usage Board. (2006, August 28). *DCMI Metadata Terms*. DCMI Recommendation. Dublin Core Metadata Initiative. Retrieved November 23, 2009, from http://dublincore.org/documents/2006/08/28/dcmi-terms/

Dublin Core Metadata Initiative. (2004, December 20). *Dublin Core Metadata Element Set, version 1.1: Reference description*. DCMI Recommendation. Dublin Core Metadata Initiative. Retrieved November 23, 2009, from http://dublincore.org/documents/2004/12/20/dces/

Eadie, M. (2008, August 7). *Images application profile*. Retrieved November 24, 2009, from the University of Bath, UKOLN Web site: http://www.ukoln.ac.uk/repositories/digirep/index/Images_Application_Profile

Farquhar, A. (2009, December). Foreword. *Planetarium*, *8*, 1. Retrieved March 9, 2010, from http://planets-project.eu/docs/newsletters/Planetarium8_December09.pdf

Fedora Development Team. (2005, October 28). *Fedora open source repository software*. Fedora Project. Retrieved from http://www.fedora.info/documents/WhitePaper/FedoraWhitePaper.pdf

Fedora Repository 3.3 documentation. (2010, January 7). Retrieved February 22, 2010, from the Fedora Commons Web site: http://fedora-commons.org/confluence/display/FCR30/Fedora+Repository+3.3+Documentation

Fedora service framework: Fedora release 2.2. (2007, January 19). Retrieved June 25, 2009, from the Fedora Project Web site: http://www.fedora.info/download/2.2/userdocs/server/features/serviceframework.htm

Feijen, M., Horstmann, W., Manghi, P., Robinson, M. & Russell, R. (2007, October). DRIVER: Building the network for accessing digital repositories across Europe. *Ariadne*, *53*. Retrieved October 21, 2009, from http://www.ariadne.ac.uk/issue53/feijen-et-al/

Ferreira, M., Baptista, A. A. & Ramalho, J. C. (2009, May 12). *CRiB: Conversion and recommendation of digital object formats*. Retrieved October 26, 2009, from the University of Minho, Department of Information Systems Web site: http://crib.dsi.uminho.pt/

Fontaine, K., Hunolt, G., Booth, A. & Banks, M. (2007, October). Observations on cost modeling and performance measurement of long-term archives. In E. Mikusch & C. Reck (Eds.), *PV 2007: Ensuring the long term preservation and value adding to scientific and technical data – conference proceedings*. Oberpfaffenhofen, Germany: German Aerospace Centre (DLR) and German Remote Sensing Data Centre (DFD). Retrieved October 21, 2009, from http://www.pv2007.dlr.de/Papers/Fontaine_CostModelObservations.pdf

Frequently asked questions. (n.d.). Retrieved February 19, 2010, from the DSpace Foundation Web site: http://www.dspace.org/faq/FAQ.html

Fry, J., Lockyer, S., Oppenheim, C., Houghton, J. & Rasmussen, B. (2008, November). *Identifying benefits arising from the curation and open sharing of research data produced by UK higher education and research institutes*. JISC. Retrieved October 13, 2009, from http://ie-repository.jisc.ac.uk/279/

Giaretta, D. (2007, June). The CASPAR approach to digital preservation. *International Journal of Digital Curation*, *2*(1). Retrieved October 21, 2009, from http://www.ijdc.net/ijdc/article/view/29/

Giaretta, D. (2009, April). Preservation workflows, strategies and infrastructure. In H. R. Tibbo, C. Hank, C. A. Lee & R. Clemens (Eds.), *Proceedings of DigCCurr 2009: Digital curation: Practice, promise and prospects* (pp. 177–184). Raleigh, NC: Lulu. University of North Carolina at Chapel Hill, NC. Retrieved October 21, 2009, from http://stores.lulu.com/DigCCurr2009

Giaretta, D., Patel, M., Rusbridge, A., Rankin, S. & McIllwrath, B. (2005, September). Supporting e-Research using Representation Information. In *Proceedings of the UK e-Science All Hands Meeting*. Nottingham. Retrieved from http://www.allhands.org.uk/2005/proceedings/papers/447.pdf

Ginsparg, P. (1994). First steps towards electronic research communication. *Computers in Physics*, *8*(4), 390–397. Retrieved February 19, 2010, from http://people.ccmr.cornell.edu/~ginsparg/blurb/blurb.ps.gz

Glick, K., Wilczek, E. & Dockins, R. (2006, September). *Analysis of Fedora's ability to support preservation activities* (Report 4.1). Medford, MA and New Haven, CT: Tufts University and Yale University. Retrieved February 22, 2010, from http://hdl.handle.net/10427/36388

Green, R. A. (2008, July 9). *RIDIR project: Final project report*. JISC. Retrieved February 15, 2010, from http://www2.hull.ac.uk/discover/PDF/RIDIR_-final-report.pdf

Green, R. A., Awre, C. L., Sherratt, R., Lamb, S. W. & Dolphin, I. (2007, October). *RepoMMan project: Final project report*. JISC. Retrieved November 19, 2009, from http://ie-repository.jisc.ac.uk/125/

Greenberg, J., White, H. C., Carrier, S. & Scherle, R. (2009). A metadata best practice for a scientific data repository. *Journal of Library Metadata*, *9*(3/4), 194–212. doi:10.1080/19386380903405090

Gregorio, J. & de hOra, B. (Eds.). (2007, October). *The Atom publishing protocol*. Request for Comments. Retrieved January 7, 2010, from the Internet Engineering Task Force Web site: http://tools.ietf.org/html/rfc5023

Hakala, J. (Ed.). (2001, October). *Using national bibliography numbers as Uniform Resource Names*. Request for Comments. Retrieved February 16, 2010, from the Internet Engineering Task Force Web site: http://tools.ietf.org/html/rfc3188

Heery, R. & Patel, M. (2000, September). Application profiles: Mixing and matching metadata schemas. *Ariadne*, *25*. Retrieved November 26, 2008, from http://www.ariadne.ac.uk/issue25/app-profiles/

Heslop, H., Davis, S. & Wilson, A. (2002, December). *An approach to the preservation of digital records*. National Archives of Australia. Retrieved January 18, 2010, from http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm2-888.pdf

Heus, P. (2008, April). *DDI-DExT Tools Project*. UK Data Archive and Open Data Foundation. Retrieved February 19, 2010, from http://www.data-archive.ac.uk/dext/reports/ODaF_Report.doc

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation*, *3*(1). Retrieved October 7, 2009, from http://www.ijdc.net/index.php/ijdc/article/view/69

Hilse, H.-W. & Kothe, J. (2006, November). *Implementing persistent identifiers: Overview of concepts, guidelines and recommendations*. London and Amsterdam: Consortium of European Research Libraries and European Commission on Preservation and

Access. Retrieved February 16, 2010, from http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8

Hitchcock, S., Brody, T., Hey, J. M. N. & Carr, L. (2005, November). Preservation for institutional repositories: Practical and invisible. In *Proceedings of PV 2005: Ensuring long-term preservation and adding value to scientific and technical data*. Royal Society of Edinburgh, Edinburgh, UK. Retrieved November 11, 2009, from http://eprints.soton.ac.uk/18774/

Hitchcock, S., Brody, T., Hey, J. M. N. & Carr, L. (2007a). Digital preservation service provider models for institutional repositories: Towards distributed services. *D-Lib Magazine*, *13*(5/6). Retrieved from http://www.dlib.org/dlib/may07/hitchcock/05hitchcock.html

Hitchcock, S., Brody, T., Hey, J. M. N. & Carr, L. (2007, January 25). *Preservation metadata for institutional repositories: Applying PREMIS*. University of Southampton. Retrieved September 29, 2009, from http://preserv.eprints.org/papers/presmeta/presmeta-paper.html

Hitchcock, S., Brody, T., Hey, J. M. N. & Carr, L. (2007, February 21). *Survey of repository preservation policy and activity*. University of Southampton. Retrieved September 16, 2009, from http://preserv.eprints.org/papers/survey/survey-results.html

Hitchcock, S., Hey, J. M., Brody, T. & Carr, L. (2007, March). *Laying the foundations for repository preservation services: Final report from the PRESERV project*. JISC. Retrieved from http://eprints.ecs.soton.ac.uk/18147/

Hitchcock, S., Tarrant, D. & Carr, L. (2009, July). *Towards repository preservation services: Final report from the JISC Preserv 2 project*. JISC. Retrieved from http://ie-repository.jisc.ac.uk/381/

Hunter, J. & Choudhury, S. (2006). PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services. *International Journal on Digital Libraries*, *6*(2), 174–183. doi:10.1007/s00799-005-0134-z

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional requirements for bibliographic records: Final report*. UBCIM Publications — New Series. Munich: K. G. Saur. Retrieved December 23, 2008, from http://www.ifla.org/VII/s13/frbr/frbr.pdf

Innocenti, P., Aitken, B., Hasan, A., Ludwig, J., Maciuvite, E., Barateiro, J. et al. (2009, March 9). *SHAMAN requirements analysis report (public version) and specification of the SHAMAN assessment framework and protocol* (Project public deliverable SHAMAN-WP1-D1.2). SHAMAN Project. Retrieved October 21, 2009, from http://shaman-ip.eu/shaman/sites/default/files/SHAMAN_D1.2_Requirements%20analysis%20report_0.pdf

Innocenti, P. & Vullo, G. (2009). Assessing the preservation of institutional repositories with DRAMBORA: case studies from the University of Glasgow. *Bollettino AIB*, *49*(2), 139–156. Retrieved February 19, 2010, from http://www.aib.it/aib/boll/2009/0902139.htm

ISO 3166-1. (2006). *Codes for the representation of names of countries and their subdivisions – Part 1: Country codes*. International Organization for Standardization.

ISO/IEC 21000-2. (2005). *Information technology – Multimedia framework (MPEG-21) – Part 2: Digital Item Declaration*. International Organization for Standardization.

Jantz, R. C., Agnew, G., Bennett, A., Bevan, P., Davis, D., Glick, K. et al. (2006). *Report from the Fedora Preservation Services Working Group*. Presentation given at the Fedora Users Conference, University of Virginia. Retrieved February 22, 2010, from http://www.lib.virginia.edu/digital/fedoraconf/presentations/fedora_2006-2_jantz.ppt

Joint Information Systems Committee. (2003, June). *JISC Circular 6/03 (revised): An invitation for expressions of interest to establish a new Digital Curation Centre for research into and support of the curation and preservation of digital data and publications*. Retrieved November 11, 2009, from http://www.jisc.ac.uk/uploaded_documents/6-03%20Circular.doc

Jones, S. (2009, October). *Data Asset Framework implementation guide*. University of Glasgow, Humanities Advanced Technology and Information Institute. Retrieved February 19, 2010, from http://www.data-audit.eu/docs/DAF_Implementation_Guide.pdf

Jones, S., Ross, S., Ruusalepp, R. & Dobreva, M. (2009, May 26). *Data Audit Framework methodology*. Version 1.8. University of Glasgow, Humanities Advanced Technology and Information Institute. Retrieved February 19, 2010, from http://www.data-audit.eu/DAF_Methodology.pdf

King, R. (2008, April 14). *The Planets Interoperability Framework*. Presented at the tutorial *Planning the Future with Planets* at the Austrian Computer Society, Vienna. Retrieved March 9, 2010, from http://www.planets-project.eu/docs/presentations/Interoperability_Framework_RK.pdf

Knight, G. (2008, June 10). *Deciding factors: Issues that influence decision-making on significant properties*. JISC. Retrieved January 7, 2010, from http://www.kcl.ac.uk/content/1/c6/04/55/43/deciding-factors.pdf

Knight, G. (2008, February 14). *Framework for the definition of significant properties*. JISC. Retrieved January 7, 2010, from http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf

Knight, G. (2008, August 21). *Significant properties data dictionary*. JISC. Retrieved January 7, 2009, from http://www.kcl.ac.uk/content/1/c6/04/55/43/sigprop-dictionary.pdf

Knight, G. (2009, April 23). *SHERPA DP2: Developing services for archiving and preservation in a distributed environment*. JISC. Retrieved November 19, 2009, from http://ie-repository.jisc.ac.uk/395/

Knight, G. & Anderson, S. (2007, March 29). *SHERPA DP: Final report of the SHERPA DP project*. JISC. Retrieved November 19, 2009, from http://www.jisc.ac.uk/media/documents/programmes/preservation/sherpa%20dp%20final%20report.doc

Kunze, J. & Rodgers, R. P. C. (Eds.). (2008, May 22). *The ARK identifier scheme*. Internet-Draft. Retrieved February 11, 2010, from the Internet Engineering Task Force Web site: http://tools.ietf.org/html/draft-kunze-ark-15

Law, K. H., Peng, J. & Demian, P. (2005, September 30). *An assessment of data curation issues for NEES* (Technical Report TR-2005-047). NEES Program. Retrieved November 11, 2009, from http://wiki.nees.org/download/attachments/689023/TR-2005-047new_Kincho.pdf

Livingston, H. & Nastasie, D. (2007, May). The role of academic libraries in the sustainability, preservation and access control of digital repositories. In *Proceedings of*

EDUCAUSE Australasia '07. Melbourne Exhibition and Convention Centre, Melbourne, Australia. Retrieved November 11, 2009, from http://www.caudit.edu.au/educauseaustralasia07/authors_papers/Livingston.pdf

Lord, P. & Macdonald, A. (2003). *E-Science curation report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision*. JISC. Retrieved November 9, 2009, from http://www.jisc.ac.uk/media/documents/programmes/preservation/e-sciencereportfinal.pdf

Lorie, R. (2002, December). *The UVC: a method for preserving digital documents – proof of concept* (Long-Term Preservation Study Report 4). IBM and Koninklijke Bibliotheek. Retrieved February 11, 2010, from http://www.kb.nl/hrd/dd/dd_onderzoek/reports/4-uvc.pdf

Lossau, N. & Peters, D. (2008, December). DRIVER: Building a sustainable infrastructure of European scientific repositories. *Liber Quarterly*, *18*(3/4). Retrieved October 22, 2009, from http://liber.library.uu.nl/publish/articles/000267/article.pdf

Lynch, C. A. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. *ARL: A Bimonthly Report*, (226). Retrieved from http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

Maniatis, P., Roussopoulos, M., Giuli, T. J., Rosenthal, D. S. H. & Baker, M. (2005, February). The LOCKSS peer-to-peer digital preservation system. *ACM Transactions on Computer Systems*, *23*(1), 2–50. doi:10.1145/1047915.1047917

Martin, D., Ankolekar, A., Burstein, M., Denker, G., Elenius, D., Hobbs, J. et al. (2004, November). *OWL-S 1.1 release*. DAML Services.

Matthews, B., Shaon, A., Bicarregui, J., Jones, C., Woodcock, J. & Conway, E. (2009). Towards a methodology for software preservation. In *Proceedings of the 6th International Conference on the Preservation of Digital Objects (iPRES 2009)* (pp. 132–140). Oakland, CA: California Digital Library. Retrieved March 9, 2010, from http://www.escholarship.org/uc/item/8089m1v1

McGuinness, D. L. & van Harmelen, F. (Eds.). (2004, February 10). *OWL Web Ontology Language overview*. W3C Recommendation. World Wide Web Consortium. Retrieved October 23, 2007, from the World Wide Web Consortium Web site: http://www.w3.org/TR/owl-features/

McLeod, R., Wheatley, P. & Ayris, P. (2006, April 7). *Lifecycle information for e-literature: Full report from the LIFE Project*. JISC. Retrieved October 13, 2008, from http://eprints.ucl.ac.uk/1854/

Mealling, M. & Denenberg, R. (Eds.). (2002, August). *Report from the Joint W3C/IETF URI Planning Interest Group: Uniform Resource Identifiers (URIs), URLs, and Uniform Resource Names (URNs): clarifications and recommendations*. Request for Comments. Retrieved February 16, 2010, from the Internet Engineering Task Force Web site: http://tools.ietf.org/html/rfc3305

METS Editorial Board. (2007, September). *Metadata Encoding and Transmission Standard: Primer and reference manual*. Version 1.6. Digital Library Federation. Retrieved February 19, 2010, from http://www.loc.gov/standards/mets/METS%20Documentation%20final%20070930%20msw.pdf

Millington, P. (2006, May). *Moving forward with the OpenDOAR directory*. Presentation given at the 8th International Conference on Current Research Information Sys-

tems, Bergen, Norway. Retrieved October 5, 2009, from http://www.opendoar.org/documents/BergenPresentation20060512Handouts.ppt

National Library of New Zealand. (2003, June). *Metadata standards framework: Preservation metadata (revised)*. Retrieved October 1, 2009, from http://www.natlib.govt.nz/downloads/metaschema-revised.pdf

Nilsson, M., Baker, T. & Johnston, P. (2008, January 14). *Singapore Framework for Dublin Core Application Profiles*. DCMI Recommended Resource. Dublin Core Metadata Initiative. Retrieved November 23, 2009, from http://dublincore.org/documents/2008/01/14/singapore-framework/

Nottingham, M. & Sayre, R. (Eds.). (2005, December). *The Atom syndication format*. Request for Comments. Retrieved February 11, 2010, from the Internet Engineering Task Force Web site: http://tools.ietf.org/html/rfc4287

OASIS eXtensible Access Control Markup Language (XACML) TC. (2009). Retrieved February 22, 2010, from the Organization for the Advancement of Structured Information Standards Web site: http://www.oasis-open.org/committees/xacml/

Ockerbloom, J. (1998, January). *Mediating among diverse data formats* (Technical Report CMU-CS-98-102). Carnegie Mellon University. Retrieved February 15, 2010, from http://web.archive.org/web/*/tom.library.upenn.edu/pubs/thesis/

OCLC/RLG Working Group on Preservation Metadata. (2002, June). *Preservation metadata and the OAIS Information Model: A metadata framework to support the preservation of digital objects*. Dublin, OH: Online Computer Library Center. Retrieved September 28, 2009, from http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

Open Archives Initiative. (2008, October 17). *Object Reuse and Exchange (ORE) user guide – primer*. Version 1.0. Retrieved September 29, 2009, from http://www.openarchives.org/ore/1.0/primer

Open Archives Initiative. (2008, December 7). *The Open Archives Initiative Protocol for Metadata Harvesting*. Version 2.0. Retrieved December 23, 2008, from the Open Archives Initiative Web site: http://www.openarchives.org/OAI/openarchivesprotocol.html

OpenDOAR. (2009, February 24). *OpenDOAR chart: Usage of open access repository software – worldwide*. Retrieved February 24, 2009, from the University of Nottingham Web site: http://www.opendoar.org/onechart.php?groupby=r.rSoftWareName&orderby=Tally%20DESC&charttype=pie&width=600&height=300&caption=Usage%20of%20Open%20Access%20Repository%20Software%20-%20Worldwide

Oppenheim, C., Probets, S. & Gadd, E. (2003, August). *Project RoMEO final report*. JISC. Retrieved February 17, 2010, from http://www.lboro.ac.uk/departments/ls/disresearch/romeo/RoMEO-Final-Report.doc

Other VTLS open source components. (2008). Retrieved February 22, 2010, from the VTLS Web site: http://www.vtls.com/opensource/other

Packager plugins. (2007, January 30). Retrieved February 19, 2010, from the DSpace Foundation Web site: http://wiki.dspace.org/index.php?title=PackagerPlugins&oldid=6698

Pearson, D. & Walker, M. (2007, November 30). *Report of the Format Notification and Obsolescence Service (AONS II)*. Australian Partnership for Sustainable Repositories.

Retrieved November 19, 2009, from http://www.apsr.edu.au/aons2/report.pdf

PILIN Team. (2007, December 20). *PILIN Project: Project closure report*. Monash University. Retrieved February 15, 2010, from http://resolver.net.au/hdl/102.100.272/RV0MMSMQH

Pinfield, S. & James, H. (2003). The digital preservation of e-prints. *D-Lib Magazine*, *9*(9). doi:10.1045/september2003-pinfield

Powell, A., Nilsson, M., Naeve, A. & Johnston, P. (2005, March 7). *DCMI Abstract Model*. DCMI Recommendation. Dublin Core Metadata Initiative. Retrieved November 23, 2009, from http://dublincore.org/documents/2005/03/07/abstract-model/

PREMIS Editorial Committee. (2008, March). *PREMIS data dictionary for preservation metadata*. Version 2.0. Washington, DC: Library of Congress. Retrieved September 28, 2009, from http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

Preservation Metadata: Implementation Strategies (PREMIS) Working Group. (2005–05, May). *Data dictionary for preservation metadata*. Final report. Version 1.0. Dublin, OH and Mountain View, CA: Online Computer Library Center and Research Libraries Group. Retrieved September 28, 2009, from http://www.oclc.org/research/projects/pmwg/premis-final.pdf

Ramalho, J. C., Ferreira, M., Faria, L., Castro, R., Barbedo, F. & Corujo, L. (2008, September 30). RODA and CRiB: A service-oriented digital repository. In *Proceedings of the 5th International Conference on Preservation of Digital Objects (iPRES 2008): Joined up and working: tools and methods for digital preservation*. London, 29-30 September 2008. Retrieved June 26, 2009, from http://hdl.handle.net/1822/8226

Reich, V. (2009, March). Distributed digital preservation. In *Proceedings of the Indo-US Workshop on International Trends in Digital Preservation*. National Digital Preservation Program. Pune, India. Retrieved November 23, 2009, from the National Digital Preservation Program Web site: http://www.lockss.org/lockysswiki/files/ReichIndiaFinal.pdf

Reid, J. (2008, March 12). *Geospatial application profile*. UKOLN, University of Bath. Retrieved November 24, 2009, from http://www.ukoln.ac.uk/repositories/digirep/images/e/ef/Geospatial_Application_Profile.doc

RLG-NARA Digital Repository Certification Task Force. (2005, August). *Audit checklist for the certification of trusted digital repositories: Draft for public comment*. Mountain View, CA: Research Libraries Group. Retrieved February 17, 2010, from http://web.archive.org/web/*/www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

RLG-NARA Digital Repository Certification Task Force. (2007). *Trustworthy repositories audit and certification: Criteria and checklist*. Chicago, IL and Dublin, OH: Center for Research Libraries and Online Computer Library Center. Retrieved February 17, 2010, from http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

RLG/OCLC Working Group on Digital Archive Attributes. (2002, May). *Trusted digital repositories: Attributes and responsibilities*. Mountain View, CA: Research Libraries Group. Retrieved February 17, 2010, from http://web.archive.org/web/*/www.rlg.org/en/pdfs/repositories.pdf

Rosenthal, C., Blekinge-Rasmussen, A., Hutař, J., McHugh, A., Witham, E. & Ross, S. (2008, March 25). *Repository planning checklist and guidance* (Deliverable D3.2). DigitalPreservationEurope. Retrieved October 5, 2009, from http://www.digitalpreservationeurope.eu/publications/reports/Repository_Planning_Checklist_and_Guidance.pdf

Sauermann, L. & Cyganiak, R. (Eds.). (2008, December 3). *Cool URIs for the Semantic Web*. W3C Interest Group Note. Retrieved February 16, 2010, from the World Wide Web Consortium Web site: http://www.w3.org/TR/2008/NOTE-cooluris-20081203/

Schmidt, R., King, R., Steeg, F., Melms, P., Jackson, A. & Wilson, C. (2009). A framework for distributed preservation workflows. In *Proceedings of the 6th International Conference on the Preservation of Digital Objects (iPRES 2009)* (pp. 162–168). Oakland, CA: California Digital Library. Retrieved March 9, 2010, from http://www.planets-project.eu/docs/papers/Schmidt_AFrameworkfor_iPres2009.pdf

Šimko, V., Máša, M. & Giaretta, D. (2009, May). Long-term digital preservation of a new media performance: 'Can we re-perform it in 100 years?' *International Preservation News*, *47*, 32–34.

Smith, C., Davis, D. & Staples, T. (2008, December 18). *Fedora Commons community registry*. Retrieved February 24, 2009, from https://fedora-commons.org/confluence/display/FCCommReg/Fedora+Commons+Community+Registry

Strodl, S. & Becker, C. (2007, July 31). *Report on methodology for specifying preservation plans* (Deliverable PP4/D1). Planets Project. Retrieved September 23, 2009, from http://www.ifs.tuwien.ac.at/dp/plato/docs/Planets_PP4-D1_Final.pdf

Tarrant, D. & Hitchcock, S. (2009, May). *The KeepIt Project: Kultur, eCrystals, EdShare (and NECTAR) – Preserve It!* Presentation given at the 4th Annual International Open Repositories Conference (OR09), EPrints User Group, Atlanta, GA. Retrieved November 20, 2009, from http://eprints.ecs.soton.ac.uk/17277/

Thomas, S. (2008, August). *Complex Archive Ingest for Repository Objects (CAIRO) project*. JISC. Retrieved September 28, 2009, from http://ie-repository.jisc.ac.uk/392/

Tonkin, E. & Strelnikov, A. (2009, April). Spinning a semantic web for metadata: Developments in the IEMSR. *Ariadne*, *59*. Retrieved July 1, 2010, from http://www.ariadne.ac.uk/issue59/tonkin-strelnikov/

UDFR Initiative. (2009, March). *Unified digital formats registry (UDFR): Proposal and roadmap*. Retrieved September 28, 2009, from http://www.udfr.org/docs/Unified_Digital_Formats_Registry.pdf

Uniform Resource Identifer (URI) schemes per RFC4395. (2010, January 13). Retrieved February 16, 2010, from the Internet Corporation for Assigned Names and Numbers Web site: http://www.iana.org/assignments/uri-schemes.html

Van de Sompel, H., Bekaert, J., Liu, X., Balakireva, L. & Schwander, T. (2005). aDORe: a modular, standards-based Digital Object Repository. *Computer Journal*, *48*(5), 514–535. doi:10.1093/comjnl/bxh114

Van de Sompel, H. & Lagoze, C. (2000). The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, *6*(2). doi:10.1045/february2000-vandesompel-oai

Vanderfeesten, M., Summann, F. & Slabbertje, M. (Eds.). (2008, November 13). *DRIVER guidelines for content providers: Exposing textual resources with OAI-PMH*. Version 2.0. DRIVER Project. Retrieved February 17, 2010, from http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf

Verdegem, R. & van der Hoeven, J. (2006, May). Emulation: To be or not to be. In S. Chapman & S. A. Stovall (Eds.), *Archiving 2006: Final program and proceedings* (pp. 56–60). Springfield, VA: Society for Imaging Science and Technology.

Watry, P. (2007). Digital preservation theory and application: Transcontinental persistentarchives testbed activity. *International Journal of Digital Curation*, *2*(2). Retrieved October 21, 2009, from http://www.ijdc.net/ijdc/article/view/43/

Woodyard-Robinson, D. (2007, June 4). *Implementing the PREMIS data dictionary: A survey of approaches*. Library of Congress. Retrieved September 29, 2009, from http://www.loc.gov/standards/premis/implementation-report-woodyard.pdf