PRESTO – Preservation Technologies for European Broadcast Archives

IST-1999-20013

D7.2 Guide to Metadata in Preservation

DOCUMENT IDENTIFIER	PRESTO-W7.2-BBC-020430 Metadata Reference Process
DATE	30 April 2002

ABSTRACT: A document will be prepared to show how metadata can be handled to reduce rather than increase total cost, and to guide archivist through the labyrinth of existing and proposed standards.

AUTHOR, COMPANY	R Wright BBC
INTERNAL REVIEWER	
WORKPACKAGE / TASK	WP7 / T2

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
1.0	17.04.2002	First draft	Draft	Confidential
2.0	09.05.2002	Full draft	Draft	Confidential
3.0	12.05.2002	Final	Final	Confidential

Table of Contents

1. Summary

There are many aspects of metadata. Two major considerations are examined in this document: cost, and standardisation.

Cost: Metadata updates in preservation work can be seen as just an extra cost, but experience has shown that an investment in metadata automation provides the central ‘process control’ for total system automation, leading to maximum workflow efficiency, high quality control and reduced cost.

Standardisation: The greatest overall benefit of investment in preservation work is obtained when the documentation of the preserved materials is in a standard format, for greatest access. Audiovisual metadata standardisation is currently a field of intense activity. Within the consortium are the key archives and personnel leading European activity relating to broadcasting, and they have links to related work in libraries and archives standardisation, and internet standardisation. There is an emerging consensus, which will be explained and justified.

This document shows how metadata can be handled to reduce rather than increase total cost, and provides a guide through the labyrinth of existing and proposed standards.

2. Metadata Savings

2.1. Overview

2.1.1. Data and Metadata

Archives are large collections, traditionally of documents but the 20th Century saw the development of media archives, especially associated with broadcasting. Large collections of anything lead to the issues of how to know what items are in the collection, and how to find those items. So from the beginning, certainly from the Alexandria library of over 2000 years ago, libraries and archives had some sort of data about the collection. In modern collections this data is usually a library catalogue. Historically a library catalogue was on cards in drawers, but for over 30 years such catalogues have been held electronically on computer systems.

So archives have the material in the archive, and they have data about the material. Any consideration of what kind of data is kept is not a question of the data itself, but of that archive's **metadata.**

An example of data: "Richard Smith, 07.12.1956, 319955B".

An example of a metadata statement: an employee's electronic personnel file shall begin with these three fields: "name, birth date, staff number".

If the BBC wants to take over another company (ACME Archives), and acquire their staff, then somebody will eventually have the task of getting the staff data from the ACME Archives personnel records into the BBC records. Name and birth date should be simple enough, though many computers and records systems differ in how they handle dates – and names for that matter. The real problem will be staff number. What information in the ACME Archives personnel data is the equivalent of the BBC staff number? That is a metadata question. ACME may have something called 'payroll code' which performs the same function as the BBC staff number. Saying that 'ACME payroll code' is equivalent to 'BBC staff number' is a statement about metadata. It is not a statement about '319955B' or 'ZXYasdf001', which are data and not metadata.

Metadata discussions have been vital in the library world for the last 40 years, as libraries tried to agree on methods to share catalogues and hence share books. Standards have been agreed (which will be briefly reviewed in the next section), and it is now commonplace for a whole network of libraries across a country or across a continent to allow a user, with one search, to find an item in any one of a large number of libraries. This is even possible for libraries that do not use identical metadata – providing there is a method for interpreting each library's metadata in a common fashion. This process is equivalent to a super-personnel system that recognises that "BBC staff number" and "ACME payroll code" are, for certain purposes, equivalent.

2.1.2. A confusion of terminology

Problems with the definition of metadata began when it became possible to combine two very different kinds of data in one electronic file or signal:

- Data directly representing the audio or visual signal (the sound or image)
- Data supplying information about the signal (identification, format, etc)

Properly, both are data. The definition of which elements to use in the "information about the signal" would be a metadata issue. But metadata would not actually appear in a file or signal, because (using the definition of metadata given in 2.1.1) metadata is only a concept and has no physical existence.

Unfortunately standard practice has now diverged from the original 1960's-style data processing definition of metadata for at least two decades, and now standard practice is to call the "information about the signal" the metadata, and call the signal itself the data. Similarly, the <meta> tag in HTML labels identifying data and not metadata per se. However we are all so far down the road in this practice that the dual usage of the term metadata has to be simply accepted. To be clear, the two usages of metadata are:

> **Abstraction**: metadata is about categories of data; their labels and arrangement.
>
> **Data:** metadata is the identifying data associated with some other data or object (web page; photograph; audio or video signal, …).

The practice of calling identifying data 'metadata' is enshrined in the original EBU-SMPTE work (REF), where **content** is defined as **essence plus metadata**, and the practice has been perpetuated ever since in virtually all broadcast-related metadata discussions.

This situation is not a huge problem, because we all manage every day to cope with terms with multiple interpretations. It is unfortunate to have one more element of confusion in an area that is already confused. It just means that broadcasting uses the term metadata in a way that can puzzle people in both the library and the IT communities. Unfortunately, the most important work on multimedia metadata involves precisely those three areas: broadcasting, IT and library science. We just have to now remember that there are two kinds of 'metadata' – the conceptual categories (the narrow definition of metadata), and the common practice where any sort of labelling or identifying text is called metadata.

Practical people should stop reading here, and skip to Section 2.2. However there is more to say on the definition of metadata. Properly, METADATA is a registered US trademark (owned by Jack Myers; coined in 1969), and we shouldn't use the word at all; we should use "meta-data" or "meta data". In practice this distinction is long buried, though the IEEE perseveres with the term meta-data: **IMS Learning Resource Meta-data Best Practices and Implementation Guide** [1]. The Worldwide Web believes that use of lowercase is distinct from the registered trademark, which it states is upper case: ""METADATA" is a trademark of the Metadata Company. W3C uses the term "metadata" in a descriptive sense, meaning "data about data". W3C is not in any way affiliated with the Metadata Company."[2]

For anyone still interested in this topic, more details of the formal definitions and history of the terms metadata, meta-data and meta data are in FOLDOC, the Free On Line Dictionary Of Computing.[3]

2.2. How metadata saves money

Section 3, below, is a summary and brief survey of the crowded area of standardisation activity with some relevance to multimedia archives. There is a lot of activity in the general area of metadata standardisation, and the whole issue can appear hopelessly complicated, with new work appearing (and old work disappearing) faster than any of us can keep up with it all, much less implement an optimum approach within our own projects and archives.

The purpose of this section is to skip to the 'bottom line' before entering the labyrinth. Metadata standardisation activity is currently complex and intense, especially for broadcast archives, which have to cope with two very different and historically unrelated areas: broadcast standards and library standards.

Despite this complexity, all archives and all preservation projects will deal with simple, basic metadata, such as item names, item numbers, and brief descriptions. There is a wide sea of various kinds of information that can be called metadata, but preservation projects have to come to grips with the simplest sort: basic item labelling.

Such metadata can, and should, be handled automatically to the greatest extent possible. If this very simple metadata is automated, manual work associated with labelling and tracking individual items is minimised and many manual steps can be completely eliminated.

Examples of manual work that can be eliminated are:

- Packing lists to accompany containers of media
- Labelling of new media
- Work lists for the daily work of preservation transfer operators
- Updating databases to show new media items replacing old

Many more examples can be given. Ultimately, ALL manual handling of metadata can be eliminated, once the initial media entering a preservation transfer process has a barcode. Using a computer to track barcodes replaces using people to do manual labelling or list making or any other basic metadata operation.

2.3. Requirements for automation:

2.3.1. Basic

Metadata automation begins with one basic step: bar coding of the physical media. This requires producing bar codes, applying them to the old media (if necessary), and applying them to the new media.

Why put bar codes on old media, if they will be thrown away at the end of the projects? Because it saves money and raises quality! It is not until a preservation transfer project is up and running that the amount of physical handling becomes evident. Work lists, packing lists, inventory, summary, check sheets – again and again everything that takes a pen and paper or a typed-in list, can be eliminated by a computer-generated list (or a computer process that doesn't even need a list) – but the material must be identified to the computer for the whole process to work. Hence the need for the bar code.

There are many basic bar code systems, and both IASA and FIAT can provide assistance if more information is needed.

Two other considerations at the basic level:

- Identifier – a standardised universal, unambiguous identifier
- Computer system

The bar code should be an arbitrary number, essentially a stock control number (acquisition number). However at this stage the bar code should link, if it doesn't already, to a standardised identifier. The principal standards used in broadcasting are mentioned in Section 3.9.

Finally, bar codes don't operate themselves. Some sort of computer and bar-code reader is needed. It should be emphasised that a basic bar-code system does not have to be expensive. The simplest can be purchased for less than 1000 Euros, complete with computer. The Google directory "Computers > Software > Bar Code" has 233 references, "Computers > Hardware > Peripherals > Bar Code" has 65, and "Business > Industries > Printing > Labels > Bar Code Labels & Equipment" has 62.

2.3.2. Intermediate

One material is bar code labelled with a computer to read the labels, the only further investment needed to actually use the computer to control the process is software. The same computer that reads and prints bar codes can, if it is a general-purpose PC. Some bar code systems use hand-held computers which are a poor option if further development is ever required.

To date, there is no general software available for running preservation transfers. This is partly because the system needs to link to the existing electronic catalogue to be efficient, and most archive catalogues are unique. It is also partly due to the lack of a common 'forward path' in archive transfers – because different archives are producing different new media.

If archives could agree on a standard for getting information in and out of their catalogues, and agree on a small range of options for production of new media, the entire preservation transfer process could be standardised. Such standardisation would allow 'preservation transfer software' to come to the market. It would also open the door for very effective use of outsourcing of the whole transfer project.

Until such standardisation takes place, individual projects can save considerable amounts of money, and raise quality, by doing bespoke programming (which can be done using 'desktop' packages such as Basic and Access) to track the media through the transfer cycle. Again, details of a generic transfer workflow are in D 7.1.

2.3.3. Full automation

For large scale transfers, not only the metadata but also the audiovisual signals themselves can be handled by the computer (once they are digitised, of course). This requires a major investment in two hardware systems:

- Central server – something that can store at least one day's work for all operations and operators, and ideally will store a week's work. This amount of storage for audio will be 50 to 500 gigabytes, and will cost somewhere in the region of 2k to 20k Euros. Video will be up to 100 times bigger, and more expensive. It may be impractical to use a video server for more than ½ day of total storage (rather than a full week), until costs drop.
- Media network – the operators and transcription stations need to be connected to the server with a high-speed network. 100 MHz Ethernet may suffice, though for video an optical network (fibrechannel) would be much better.

A chief advantage of sever-based working is for efficient handling of 'bits and pieces'. Material on many short tapes, or multiple sides of gramophone discs, can be most conveniently assembled on a server. With the computer holding both data and metadata, complicated and error-prone manual processes can be eliminated, with the computer essentially doing all processes beyond the initial digitisation.

Automated quality control: Finally, with the signal on a server, it is possible to allow the computer to check the signal as well as checking everything else. The signal can be computer-monitored at the point of digitisation, allowing operators to due more simultaneous transfers without reducing quality. The other processes that can be implemented once a server is used are described in Section 2.6.1.

2.3.4. High-volume automation

For the largest projects, there is further investment for the ultimate in automation: automatic handling of physical media.

Rather than copying “old tapes to new tapes”, server-based digitisation allows the new media to be completely automatically generated. If datatape is used, this means a **tape robot** will be used to move tapes in and out of the write mechanism.

Even for project which do not make datatape, there are full-automation possibilities. Both CDs and DVDs can be written, and labelled, by automatic equipment. Several audio archive projects now use CD writing and labelling equipment.

2.4. Benefits of automation

Cost, Quality and Time: It is generally considered that projects have parameters of cost, quality and time, and that any one can be reduced only if one or both of the others are increased. This is not the full picture, because it does not include overall volume.

An investment in automation allows cost and time to be reduced without sacrificing quality. A hand-made care would be very expensive, easily 100 times the cost of a mass-produced car, for equivalent quality – at taking a much longer time. The whole issue in using automation to break out of the cost, quality and time triangle, is volume. The automation investment has to be paid off as so much extra cost per item. Once the automation cost is below the manual cost, automation pays.

Manual throughput: The state of the art in preservation transfers, when not using automation aids, is approximately two hours of operator time. Typically an operator does a single transfer at one time, or at most two – and spends a lot of time with equipment set-up and signal monitoring.

Full automation: Those projects that have used automation can do at least four times better: approximately 1/2 hour of operator time per hour of audio (from 6mm tape) for the RAI and BBC projects (with higher throughput at RAI because of automated quality control).

Return on investment: The rest of the calculation of the benefit of automation is straightforward. If labour is reduced by 75%, and if 100,000 hours of material are involved, then there is a savings of 150,000 hours (because the manual process uses two hours of labour for hour of material). The monetary savings is company labour rate, times 150,000. That is likely to be in the region of 2 million Euros for European labour rates. An automation investment of 1 million Euros is thus very cost effective, even for ‘just’ 100,000 hours.

In short, automation saves 1.5 hours of labour costs per hour of material, which is very roughly 5 to 10 Euros of savings **per hour of material**

The RAI project was 300,000 hours, and other broadcasters will be pursuing projects of at least 100,000 hours. A fully-automated audio preservation facility such as BBC Maida Vale has a cost of roughly 250k Euros, so it pays for itself on 25k hours of material.

2.5. Preservation, Automation and Standardisation

This section is basically a guide to the maze of standards. We attempt to show which standards are relevant, and how the standards fit together. We begin with the two essentials: **identifiers**, and **basic archive descriptive metadata.**

The remaining sections are about the use and management of metadata: its physical form (**expression**), and how to **store** and move (**exchange**) metadata.

2.5.1. Identifiers

It is essential to give media an identifying number that is consistent across a collection, and ideally is unique: no other item, anywhere, with the same number.

An acquisition number will satisfy the consistency issue, and be unique across a collection. In order to be globally unique, it is necessary to follow an international standard.

There are two main forms of identifier of interest to multimedia preservation: universal identification numbers, and universal resource locators.

Universal identification numbers are based on the pattern of the ISBN, the International Standard Book Number. Two are of immediate concern:

USID: **EBU Recommendation R99-1999:** 'Unique' Source Identifier (USID) for use in the OriginatorReference field of the Broadcast Wave Format[4]

UMID: SMPTE Universal Media Identifier (SMPTE 330M-2000 Television - Unique Material Identifier (UMID))[5]. This is more complex than the USID, and includes provision for version and 'instance' control. These complexities have proved difficult to implement, and the standard is at present (May 2002) under review.

Universal resource locators refer to web technology, and include not just the URL but related URI (indicator) and URN (name). Although it is attractive to think of using this technology for identification purposes, there are two difficulties:

Registration: URL's and their relations have to be approved and released for use, which takes both time and money (though URLs can be registered for less than 100 Euros).

Persistence: The Internet is rapidly developing, and so use of a web locater as an identifier is probably not the safest and simplest method of permanent numbering. There is a widespread Internet problem with disappearance of URLs as websites come and go. Even if a company ensures that their website is not lost, it may well be that the URL for the company will need to change (change of ownership or policy), making a problem for all material which uses the URL for identification.

Further reference: Unique Identifiers: a brief introduction; Brian Green and Mark Bide[6]

2.5.2. 'Standard metadata'

In the archive world, standard metadata is shorthand for standard or core bibliographical data, or core records. It is the term for the set of information in the catalogue. Sometimes this concept is divided into **core** and **full** metadata, and both have their place. In the standardisation world, however, core or minimum metadata is by far the easiest to standardise. In consequence, there is very general acceptance and implementation of Dublin Core metadata (http://uk.dublincore.org/, see Section 3.3.3), whereas larger sets remain contentious.

The three fuller sets of metadata of immediate relevance to audiovisual archives are those proposed by IASA, FIAT and the EBU. All three are described in Section 3 of this paper.

The important question when deciding about metadata standardisation is: what do you want to do. Will your metadata stand alone, or will it be combined with similar metadata from other companies? If you are selling footage and want to sell in some general electronic marketplace, then you need to adopt the standard of that marketplace. Unfortunately we have at present more planned than actual e-markets, but the principle is clear.

If your metadata will stand alone, then it is not imperative to use anybody else's standard, though this approach (which has been commonplace in specialist archives, including broadcasting) is not only short-sighted but is very likely to become a real problem in a world where networking and interchange are dominant activities.

It is important to realise that simple, well-defined and well-managed metadata will usually be able to be **mapped** from a local system into a general system. It the time when your stand-alone archives enters some cooperative or commercial activity, data transfer or data interpretation will be required. As part of this exercise, your non-standard metadata can be mapped into a standard. This can be done relatively easily even for very large catalogues. The BBC, INA and ORF catalogues were exported, mapped to Dublin Core elements, and formed into a **union catalogue** as a relatively small part of the PRESTO project (D7.3: Common access to broadcasts archives).

2.5.3. Expression and exchange of metadata

What does metadata look like? It comes up as words on computer screens. What else do we need to know?

The problem is not how people see metadata, but how computers see it. In order to exchange data between companies, or form union catalogues, or sell to a general e-market, data has to move from computer to computer.

Computers are very rigid and only communicate when there are no uncertainties. Either two computers will hold metadata elements in exactly the same way (unlikely), or there will be a special transfer operation involving labelling the metadata at the export from one computer, to guide the interpretation at the receiving end.

Such labelled data is easily handled with a mark-up language, SGML, HTML and now XML. A mark-up language adds information to identify the data being passed. This if one computer holds **Title** as a database field of a certain length and in a certain character set, and the other has a database field called **Designator** which is the equivalent, but differs in name, length and character set – a transfer can still be made using XML. A computer programme needs to be written to look for items marked **Title**, convert them to the storage requirements of the receiving computer, and store them in the field **Designator**.

In XML there are further structures such a schemas to allow the computers to know that the data coming in is correct and complete, and guide the conversion into the receiving database.

The important issue is that metadata will (and should) be stored in whatever way a particular database prefers. This in no way invalidates the ability of that metadata to be in agreement with a standard, because standards are about the interpretation of the data, not the storage. When metadata is brought out of the database it has to be in some physical format (ie expressed), and XML provides a very attractive method of expression of the metadata.

2.5.4. Storage of metadata

In Section 2.5.3, it was assumed that metadata was in a database. However we have also discussed metadata accompanying data in an electronic signal or file. This file can be permanently (we hope) held on physical media such as datatape, CD or DVD. So why not put metadata with the media, rather than in a database.

The short answer is that metadata held with the media is on **no use at all** for helping to find material in a collection. The whole point is to have a catalogue, index, inventory, finding aid or whatever you call it – available in some way that is separate from the collection, in order to find material in the collection. Therefore the metadata must have a separate home.

It may reside with the media as well, as a sort of distributed security copy, but that raises the problem of data update. Databases are easy to update: that's what they do. Metadata written to a CD is impossible to update.

The general conclusion is: only one item of metadata should always be with the media if held in a file format: the universal identifier. That is the key to the database. Full metadata should always and only be held in the database. Other core metadata, providing it will not change, can also be held in the file, and this is the approach used in the Broadcast Wave Format.

It is to be hoped that the file formats under development for video (MXF) and film will adhere to this principle. PRESTO D5.3: A generic high-resolution format for digital film, with efficient multi-format conversion is a file format that does avoid loading down the standard (and the resultant files) with unwanted, unneeded and problematic embedded metadata.

2.6. Metadata and Preservation

There are three areas of metadata that deserve particular attention during preservation work:

- Preservation metadata – metadata specifically about the transfer process; this topic is covered in Section 3.2
- Creation of new metadata – there is a once-in-a-lifetime opportunity (in the lifetime of the audiovisual signal) to use the preservation transfer (the digitisation) to compute numerical metadata about the signal. More information is given below.
- Problems with old metadata – preservation work exposes the gaps in existing (legacy) metadata. This issue can have significant extra cost, as discussed below.

2.6.1. Creation of new metadata

An important aspect emerging from the PRESTO user survey (D2: Archive preservation and exploitation requirements)[7] is that preservation of the archives is not only aimed at preserving them for the future, trying to limit as much as possible the loss of valuable information stored within the archives on supports getting obsolete, but also at enhancing the possibility of reusing archive items in new productions. This view emphasises the service and commercial opportunities associated with preservation work. However it requires appropriate technology to maximise access (to the right material; to the greatest number of users; with the greatest speed and economy).

It is of course desirable to collect and store all available metadata, in a way compliant to the most recent and useful standards. But there is also an opportunity to create new metadata. In particular, digital signal processing techniques can be used to categorise audio/visual content (fast/slow paced footage, music/speech/silence intervals) and to summarise it, by selecting the most salient frames, so that compact catalogues can be automatically built and published on-line.

Research in the field of automatic analysis of audio/visual material is now reasonably mature, and beginning to be successfully applied in real situations[8]. This statement is particularly true for speech recognition, which is now implemented at least on a pilot basis in several archives, including the BBC.

There is a standardisation issue here. If detailed processing is undertaken, will it be future-proof, or will there be much better processing available in a few years? The short answer, in the author's view, is that audio and image processing is NOT future proof. Some protection will become available through further development of MPEG-7, which at least makes a common standard for the expression of 'numerical metadata'. But this area remains highly proprietary.

Speech processing is currently at a level (60 to 70% correct recognition of words) where the resultant data, though unusable as a transcript, is still a rich resource as a 'finding aid': free-text searches on the data works effectively, as shown by various research projects (Thistle[9], Olive[10]) and also by the personal experience of the BBC Sound Archive.

2.6.2. Problems with old metadata

Everyone beginning preservation work should be prepared to invest appropriately in cataloguing effort. It will be an unbalanced outcome of the project, if the signals (the audio and video material itself) are brought up to the latest technology, and the catalogue is not also updated.

It is inevitable, in our experience, that there are gaps in the existing metadata. These gaps will require research and manual effort to plug. The issue may be data still on cards or microfiche, rather than in electronic form. It is also common to find that there is little or no metadata, beyond information on tape boxes. In a broadcast archive, information can be found from a variety of sources (listings magazines; reference works; even websites) if the material has been broadcast or relates to a programme that has been broadcast. However converting this information to a catalogue is a slow task, taking anything from one to ten hours per hour of material, depending upon the detail of the indexing.

Allowance needs to be made for this metadata work when setting up a preservation transfer project and budget. It is not uncommon for this cost to be fully half of the total project cost.

The final issue about legacy metadata is finding a way to use the information during a preservation project, with minimum manual handling. This topic is dealt with in more detail in PRESTO D7.1: Metadata Reference Process. The problem is getting the data from the electronic catalogue into the process-control system of the transfer process, and then getting the updated data back into the catalogue. All projects known to PRESTO have had to do custom programming in this area. The development of XML (see Section 2.5.3) as a standard for transport of metadata makes these transfers easier, but again this is an area where money must be spent to achieve automation.

3. Standards

There are two sorts of standards: common practice, and formal standards defined by the various standards organisations. It is important to have a good understanding of common practice in order to put formal standards into perspective.

Very good general introductions to metadata are available on the web. Two that are recommended are:

The Getty Research Institute: Introduction to Metadata[11]

On the official National Library of Australia site: Meta Matters: What is Metadata?[12]

The above introductions give a good general picture of what metadata is and how it is used, though neither says much about audiovisual material or broadcasting. It is much less easy to give a general introduction to formal standards, because of the many kinds of metadata, and the many uses. The next section gives broad categories of metadata (though there is no standard for metadata categories, which is a meta-metadata issue), and then the following sections try to guide the reader through the metadata maze by treating individual topics one at a time.

3.1. Kinds of metadata

There are various ways to divide metadata according to its function. One method is presented in the Getty Introduction to Metadata. It is too easy to think of metadata as just descriptive (for librarians), or just technical (for engineers) or just administrative (for computer programmers). Five broad types of metadata are: administrative, descriptive, preservation, use, and technical. The following table is taken from the Getty Introduction[13]

Administrative	Metadata used in managing and administering information resources	- Acquisition information - Rights and reproduction tracking - Documentation of legal access requirements - Location information - Selection criteria for digitisation - Version control and differentiation between similar information objects - Audit trails created by record keeping systems
Descriptive	Metadata used to describe or identify information resources	- Cataloguing records - Finding aids - Specialized indexes - Hyperlinked relationships between resources - Annotations by users - Metadata for record keeping systems generated by records creators
Preservation	Metadata related to the preservation management of information resources	- Documentation of physical condition of resources - Documentation of actions taken to preserve physical and digital versions of resources, e.g.,

		data refreshing and migration
Technical	Metadata related to how a system functions or metadata behave	- Hardware and software documentation - Digitisation information, e.g., formats, compression ratios, scaling routines - Tracking of system response times - Authentication and security data, e.g., encryption keys, passwords
Use	Metadata related to the level and type of use of information resources	- Exhibit records - Use and user tracking - Content re-use and multi-versioning information

3.2. "Preservation metadata"

The transfer from an analogue to a digital carrier is a unique moment in the history of a signal. Ever afterwards, preservation transfers should be much simpler and cheaper, for the following reasons:

- perfect copies can in principle always be made.
- digital carriers can often be handled automatically
- skilled operator intervention should not be needed to make a perfect copy
- status of the media can be automatically checked on a regular basis

The transfer from analogue is when quality will be lost, if not properly managed. It is also when the total signal will be monitored by both people and equipment, and any problems or defects can be logged. This log, the 'preservation metadata', is a vital document. It says how good the original was, and gives specific information on any problems.

This information has two uses. First, is says whether an item is of adequate quality to be used as is, or whether it has problems requiring restoration. Second, if restoration is required, an operator will not have to go through the material second-by-second, because all the problems will be noted in the log. Thus the log can reduce the cost of subsequent restoration activity, even years later.

Technical operators have always made notes during media handling. The significance of 'preservation metadata' is to keep these notes in a standard format (machine readable), and preserve them as carefully as the media itself is being preserved.

In the following examples, we list:

- the simple preservation metadata collected by the BBC in the 2000-2001 Radio One preservation project
- the full metadata developed by the PRESTO project
- the EBU standard for audio preservation metadata

Example 1: Preservation Metadata from BBC audio preservation

A full description of the BBC "Radio One Archive" preservation project has been published by the Audio Engineering Society[14], and is available from the BBC

The basic metadata is written to the "Broadcast Extension" of the EBU-Standard Broadcast Wave Format (BWF) file. This standard is described in Section 3.7.

The control computer holds the full information for the BWF standard, which is gathered automatically during the Radio One preservation process, except for the Description field. The following fields are of special interest:

FieldName	Description
Org Id	BackReference to ID of original material (not stored in WAV)
Version ID	BackReference to ID of original material (not stored in WAV)
WavName	Original Filename of WAV
Description	(See below)
USID	Universal Sound Identifier
UMID	Universal Media Identifier
ClipCreate DateTime	Creation date and time

A strict format, defined by the BBC, is used for the Description field to support future machine processing and interpretation:

A=Author or Artist
D=Original Recording Date
F=Filename
K=Keywords
L=Length (duration)
P=Programme Title
R=Reference (source identifiers)
T=Title
V=Version information

e.g.
A=Crosby, Bing
D=01.03.1948
F=MV234456.WAV
K=popular; Christmas;
T=White Christmas

Finally, technical notes (by the operator) cannot be stored effectively in the BWF file itself, but they are collected and stored, in a standard format, in a separate TEXT file on the DVD that is the file-format output of the BBC process.

Example 2: Preservation Metadata from PRESTO for audio preservation

The PRESTO project has developed a full automation system for audio transcription. This is fully described in PRESTO Deliverables **D4.1 Audio quality monitor** and **D4.2 Process analysis for audio.** The following two tables give the metadata automatically collected during the transfer process. This metadata is similar that required by the EBU Standard 3285 – Supplement 2 (Capturing Report) described in Example 3, below (and in Section 3.7).

Metadata generated by the Audio Chain: Housekeeping

Element	*Content*	*Description*
Hostname	PRESTO	PC hostname that runs the RealTimeStation
Operator Name	John Smith	Operator who started the transcription
Media Id	TESTID_0000000000	Identifier associated with analog media
Title	Die Zauberflote - act 1	Title associated with analog media
Author	W.A.Mozart	Media author
Recording Start	08/05/2002 12:32:47	Date and time which the recording started at
Recording End	08/05/2002 12:34:05	Date and time which the recording ended at
Duration	00:01:18	Total recording duration
Media type	TAPE	TAPE, VINYL, OTHER
Media Controller	STUDER	STUDER, TURNTABLE, MANUAL
Recording End Mode	Manual	Manual or Automatic Stop

Metadata generated by the Audio Chain: Quality Control

- ***azimuth degrees***: the value represents the time shift between left and right channels. It is given in degrees, with reference to time shifts between two 10KHz tones: since the sample frequency is 48KHz, 1-sample-shift corresponds to 75 degrees.
- ***sample frequency***: the sample frequency of the input audio signal. Actually, the monitors works only with signlas sampled at 48KHz.
- ***bandwidth***: for each channel, it is the estimated bandwidth of the input audio stream.
- ***dynamic***: for each channel, the value (in dB) represents the difference between the max peak and the noise floor. The energy of the noise is estimated automatically.
- ***SNR***: for each channel, the value (in dB) represents the signal to noise ratio, that is the difference between the energy of the audio signal and the noise floor. As in the case of dynamic, the energy of the noise is estimated automatically.
- ***DC-offset***: for each channel, it represents the value of the direct current component in the input audio stream. It is estimated through the mean of the sample values.
- ***peak***: for each channel, it is the maximum value (in dB) of the samples.
- ***energy***: for each channel, the value (in dB) represents the energy of the input audio stream. It is estimated through the mean square amplitude.
- ***clicks***: for each channel, the value represents the number of clicks detected per million of samples.
- ***silence***: (only off-line monitor) for each channel, the value represents the percentage of silence, or better noise floor, detected. Time intervals where silence occur are specified in the segmentation section (see below).

- ***saturation***: (only off-line monitor) for each channel, the value represents the percentage of saturation detected. Time intervals where saturation occur are specified in the segmentation section (see below).
- ***segmentation***: for each channel, the section contains the time segmentation and classification of the audio stream; homogeneous (from the acoustic viewpoint) segments are classified in terms of the following audio classes:
    - silence
    - noise
    - music
    - speech:
        - gender: female/male
        - bandwidth: narrow/wide

In each audio segment, saturation intervals (if any) are also specified as subsegments.

Example 3: EBU standard for preservation metadata for audio

As described in Section 3.7, the EBU has standards in audio relating to preservation transfers. In this section we list (in a shortened form; full details in the standard) the elements of the **Quality Chunk** (EBU Standard 3285 – Supplement 2 - Capturing Report [15]), for comparison with the two previous examples:

<< Basic data >>

Archive no. (**AN**):
Title (**TT**):
Duration (**TD**):
Operator (**OP**):.
Copying station (**CS**)

<< Start modulation data >>

SM= starting time of the sound signal from the start of the file.
Sample count (**SC**): from the start of the file
Comment (**T**)

<< Quality event data >>

Event number (**M**): Numbered mark originated manually Format: << Event number (**A**): Numbered mark originated automatically
Priority (**PRI**): Priority of the quality event
Time stamp (**TS**):
Event type (**E**): e.g. .Click., .AnalogOver., .Transparency.
Or QualityParameters (defined below) exceeding limits,
e.g. .QP:Azimuth:L-20.9smp..
Status (**S**): e.g. .unclear., .checked., .restored., .deleted..
Comment (**T**)
Sample count (**SC**):

<< Quality parameter data >>

Parameters (**QP**):
MaxPeak
MeanLevel:
Correlation:
Dynamic range:
ClippedSamples:
SNR: (Signal-to-noise-ratio)

CAPTURING REPORT

Bandwidth:
Azimuth:
Balance:
DC-Offset:
Speech:
Stereo:
Quality factor (**QF**):
Summary quality factor of the sound file [1......5 (best), 0 = undefined]
Inspector (**IN**): name of the person inspecting the sound file
File status (**FS**): Ready for transmission?
[Yes / No / File is ready / not ready / undefined]
Operator comments **T=**

<< End modulation data >>

EM= time of end of modulation of the sound signal.

<< Cue sheet data >>

Cue number (**N**): Number of cue point automatically originated by the system.
Time stamp (**TS**): the time stamp of the cue point.
Text (**T**): comments of the cue point
e.g. .Beginning of an aria..
Sample count (**SC**): Sample address code of the TS point (hexadecimal ASCII)

3.3. Library standards

Library science and standards have a long history – at least 200 years. Because library standards are essentially about cataloguing in all its aspects, all this work is in some way about metadata.

There are library standards and technology in the following broad areas:

Indexing

Controlled vocabulary

Classification

Computer systems

Organisation of electronic catalogues

Access to electronic catalogues

Essential metadata: Dublin Core

More Information about library technology and standards can be found on the website of IFLA, the International Federation of Libraries and Archives:

http://www.ifla.org/II/metadata.htm

3.3.1. Indexing

Indexing is about use of metadata in such a way that the contents of a collection can be efficiently located, by a manual method (card file) or by a computer. Works of fiction in a conventional library may be arranged on the shelves by author, but they can still be found by title, through the use of a title index. Librarians thus developed data models and technology to use data fields more than 100 years before the modern era of data processing. Typical types of index field are: title, author, subject, media type, date (chronological index), location.

Controlled vocabulary: because index material is meant to be a primary 'way in' to a collection, it is important to be consistent, and not to make mistakes. Spelling words in various ways, or using a variety of near-equivalent terms, destroy the integrity and utility of an index. It is hopeless if the user has to try all sort of different searches, when a single search would suffice if the indexing were consistent. Therefore vocabulary control is essential for quality description and retrieval. Standards exist for controlled vocabulary in various areas: subject headings, country names, person names, date, location, media and many more. Where a standard does not exist, a library will still try to have a consistent source of terminology – an **authority.** The use of authority structures, which in a computer system can be simple pull-down lists, is the simplest and most effective way to control and upgrade the quality of indexing.

Allowing uncontrolled input of descriptive terms (uncontrolled keywords and phrases) is where most untrained people start when implementing metadata. It seems sensible, but is actually an incredibly poor way to build indexing for a large collection. Either material will be lost, or a huge effort will have to be put into finding ways to retrieve information despite variations in the terminology (and spelling) of the uncontrolled indexing.

Classification: A controlled vocabulary is, still, just a list. A person using a keyword index, looking for a keyword not in the list, is in deep trouble. " *Cougar* isn't in the list, now what do I do?" If the person is very lucky, the list includes terms that aren't the **preferred term,** and the user will be told "*Cougar:* see *Puma*". However there is an enormous step upward from a simple list which can be implemented, and that is to organise all the index terms in an overall structure. Thus "animal –> mammal –> large cats –> cougar (see Puma)" provides a classification structure to prevent the user being hopelessly lost. If cougar isn't there, a quick track down the chain "animal –> mammal –> large cats" will get the user into the general area. This aspect of library science has been rediscovered in recent years by the information science world (under the names 'taxonomy' and 'ontology') and is undergoing a revival of interest and computer-system support.

The ultimate in control, specificity and language independence is to code the classification system numerically. Numerical classification systems such as Dewey and UDC are the ultimate weapon of library science. A primary area of current research is in methods to combine the most powerful developments in non-library retrieval (free-text search; web crawling; content-based retrieval) with conventional indexing and classification in order to best combine the strengths of both.

3.3.2. Computer systems

Libraries have been using computer systems for 50 years, and have a long tradition of standardisation of library metadata.

Organisation of electronic catalogues: The original standard for electronic library metadata dates from around 1970, and is MARC: Machine Readable Cataloging[16]. This standard allowed libraries to implement catalogues that could be shared, read by a common interface (very useful when the web was developed), and most importantly would allow data to migrate from old to new systems, and from vendor to vendor.

MARC remains the dominant standard for the low-level organisation of bibliographical data, though the need for such 'internal standardisation' is now much less that when MARC was developed. Current IT technology is much more adept at 'metadata mapping', allowing metadata equivalence to be implemented at a high level, or at the time of data import / export / exchange, rather than being imbedded in the basic data model of the electronic catalogue.

Access to electronic catalogues: MARC is a standard at the data representation level. It allowed data to migrate from MARC-system to MARC-system, preventing catalogue obsolescence. The development of wide-area networks led to a demand for a standard at the level of functionality, so that one system would operate the same as another when connected together. In particular, librarians wanted to issue one query to many systems, and get a standard reply from all of them.

This functionality has been developed for library systems, with the unromantic name of Z39.50. It goes far beyond the database compatibility established in the IT world with SQL (see Section **Error! Reference source not found.**). You need to know the structure of a database before writing an SQL command. Z39.50 has standard 'library function' commands (like title or author search) that will operate on all Z39.50 systems.

It is unfortunate that, in broadcasting, the dominant IT effort has nothing to do with archiving, and Z39.50 functionality is unknown. Most major broadcasters are now beginning a path of company-wide metadata management (whatever they call it, that's what they're doing) at a point which library systems reached 20 years ago. The problem is that the general requirements of broadcasting do not fall into the functions implemented for library systems, so Z39.50 will not serve as a general model. The result is that archives within broadcasting will have to implement company-wide primitive system- and data-compatiblity without use of Z39.50, whereas archives which are independent have a much cheaper, more powerful and proven method for the standardisation of all functions involving archive metadata, simply through the use of Z39.50[17].

A very common type of search and retrieval systems in the library world is the systems, which are based on the Z39.50 communication protocol[18]. This protocol provides the means for generic search and present operations. It is message oriented and the most important message types are as follows:

- Init

  For setting up a session.

- Search

  Searching in a catalogue and creating result sets for later retrieval.

- Present

  Retrieving records from previously created result sets with the flexibility of requesting different record syntaxes, element sets…

- Scan

  Browsing an index (e.g. creator, date etc., as a starting point in newly queried catalogues).

Via the so-called extended services, inter library loan functionality can be made available. To assist Z39.50 clients in accessing Z39.50 servers an additional service provides the means for automatically configuring the clients: the so-called Explain service. A Z39.50 based system may look like the one in Figure 70. Different client programs on various platforms can establish Z39.50 sessions with the Z39.50 target server system (via connections utilising e.g. TCP/IP). The client can then submit search and retrieval requests to the server. The generic front-end system of the target server translates the requests (queries) into the native interface formats (e.g. SQL) of the OPAC databases.

The generic messages are based on attribute sets, which are defined for/by the different communities communicating with the Z39.50 protocol. Internationally agreed application profiles are also defined on the existing standard. These profiles include information on the offered services (i.e. the allowed message types), the involved attribute (sets) and the record formats to be used within the exchanged messages. Example profiles are: the BATH profile (for Library Applications and Resource Discovery), the CIMI profile (for Cultural Heritage Information), the Union Catalogue profile (National Library of Australia), etc.

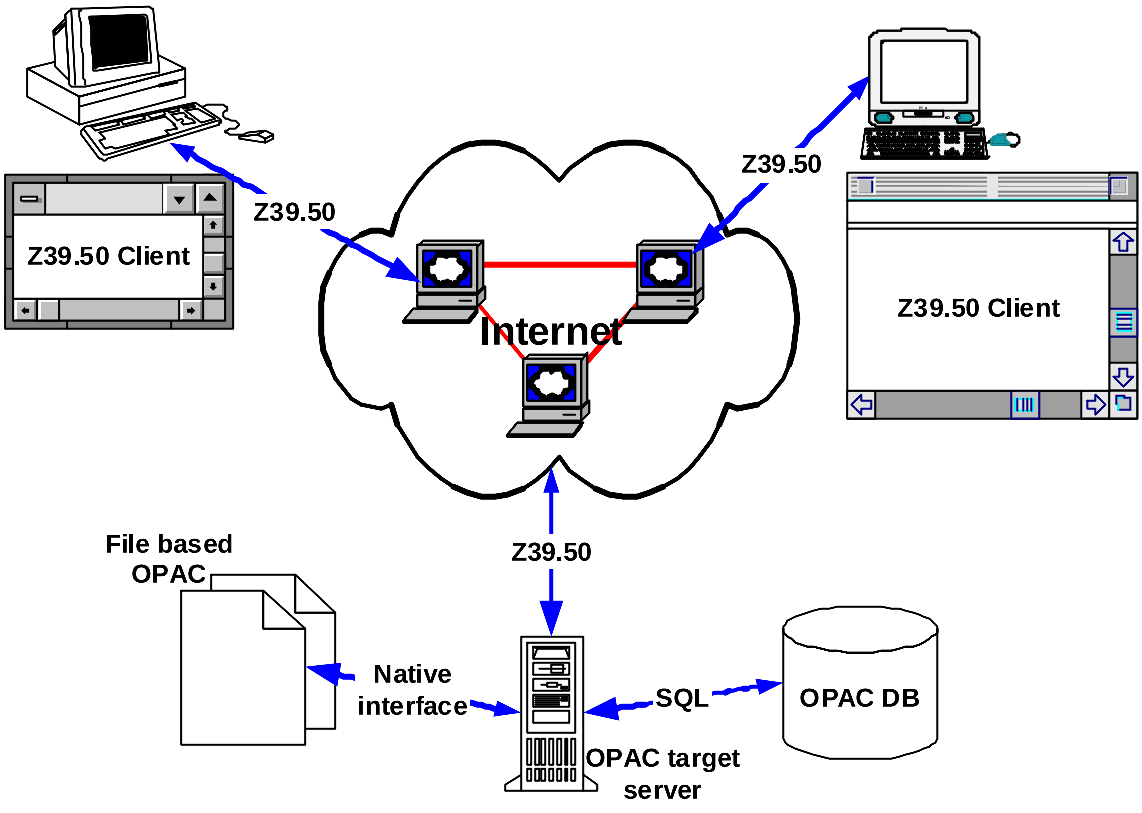


Figure 1: Z39.50 based client - server model

A wide range of applications exists in that area. Most of them are attached to existing cataloguing systems. Within these systems the OPAC database is either part of the production system or exists in parallel to the production system.

Providing access to a Z39.50 based target system via a firewall usually means that one port as to be opened (i.e. the standard Z39.50 port 210). This allows clients to connect to the server from outside the firewall. For each session one port will be opened by the server to the outside world, which is valid in many cases. There also exist web gateways (HTTP to Z39.50 and vice versa). Such gateways also allow a public access – the HTTP port is very often opened. One drawback of such solutions is the inefficiency of the mechanism. In the current situation the Z39.50 community discusses the use of alternative approaches like a representation of the messages in XML or the use of SOAP as the means of message transport.

3.3.3. Dublin Core

The single most common metadata standard is Dublin Core[19].

Dublin, Ohio, USA is the home of OCLC (Online Computer Library Centre). The Dublin Core 15 Element Set was proposed and published as DC version 1.0 in December 1996 by the Dublin Core Metadata community.

The Dublin Core Metadata Element Set (DCMES) grew out of a recognized need for improved discovery of web resources. Initially it focused on the requirement of simplicity: “ordinary” users should be able to formulate descriptive records based on a relatively simple scheme. But over the years there has been a movement to use the DCMES for more complex and specialized resource description tasks and, correspondingly, to develop mechanisms for incorporating such complexity within the basic element set.

This work is called *qualified Dublin Core*. There is a consensus, which began with the community of “web resources” (and includes library and archive communities), that Dublin Core is a suitable general approach for the standardization of metadata. Dublin Core is now a US NISO standard (Z39.85) and ratification by ISO (TC 46) and CEN is in progress. It has obtained increasing support since it was consolidated in 1996 and it is obvious that it has many qualities:

- it is a relatively simple format that can be extended, without limit, with local fields;
- it has international support;
- it has proved helpful to users in finding things;
- it is widely recognized and supported;
- it can be used directly in websites and as records in a database because of the way it is structured;
- it is maintained in a stable environment;
- its continuing development seems assured.

And, it is proving to be hospitable to a wide range of disciplines and domains, including sound recordings and moving images.

Dublin Core has been used in standards directly relevant to preservation of audiovisual media, as developed by the EBU[20], AES (Audio Engineering Society), OAIS[21], CIMI[22] and Cedars[23]. Dublin Core itself is being standardised at the international level by ISO, and at a European level by CEN.

3.4. Open Archive standard – OAIS,

Broadcast archives link to other archives which link to libraries which link to digital libraries. It is perhaps a long chain, but digital library technology in one form or another is always relevant to the future development of broadcast archives.

Broadcasters have a long commitment to open standards, in order not to be dependent upon particular manufacturers and suppliers. The same interest governs the development of digital libraries and there information systems. At the advanced end of the library community, as represented by the RLG (Research Libraries Group), there is a complex project to develop and standardise the whole future of information access in digital libraries. This effort is OAIS.

It may not be of immediate practical significance, but broadcast archives that work with national archives will already have contact with RLG or its technology.

Knowledge of OAIS is not essential to do preservation transfers, but the long-term strategy for metadata development should include a sufficiently wide horizon to take in this major effort at the top end of the standard library world.

The particular significance of OAIS is that it is concerned with preserving metadata. We in broadcasting and related archives have such a difficult situation preserving the audio and video, that we may overlook the fact that metadata itself needs preserving. The work of the OAIS provides high-level guidance, to libraries and especially to the library technology industry, on the issue of "preserving and maintaining access to digital information over the long term".[24]

3.5. Media Archive standards: FIAT, IASA

The two main international bodies for audio and television archive are IASA and FIAT.

FIAT is the International Federation of Television Archives, and IASA is the International Association of Sound and Audiovisual Archives.

Both have standards and recommendations that are of use and interest.

IASA: For general guidance in preservation work, IASA have a general guidance document (IASA TC-03): The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy[25]

For metadata, IASA have a very comprehensive document on cataloguing rules: [26] THE IASA CATALOGUING RULES for audiovisual media with emphasis on sound recordings.

FIAT: A standard, which is more focused on audio-visual archives is the Minimum Data List (MDL) provided by the FIAT/IFTA. It is used for cataloguing video and film documents. It specifies a core set of data field descriptors, which allow the set-up of new cataloguing systems in broadcast archives but also the communication of metadata between such archives. The MDL consists of three groups of descriptors: one for purpose of identification (title, date, producer…), one for describing technical data (keywords, content, format, language…) and one for data related to the management of rights (origin of the material, contracts, copyrights…). The MDL does not yet exist in a machine-readable form.

3.6. Expression and Description – XML and MPEG-7

XML[27] is increasingly the preferred method of "expression" of metadata, when metadata has to move from system to system, as discussed in section 2.5.3. MPEG-7 is a standard for the *description* of metadata•, and usually uses XML for the *expression* of that description. (All rather detailed and tedious, but you only have to read it. Think what it must be like to write.) MPEG-7 could, eventually, become important to broadcast archives, if two things happen:

• Yes, that makes it a meta-metadata standard; if we weren't all misusing the term *metadata,* and used the term *descriptive data,* we wouldn't end up with meta-metadata.

- There is sufficient agreement on and stability for the feature extraction (the calculation phase that could be performed during preservation transfers) to make these calculations sufficiently future-proof to be worth the effort.
- There is growth in the automation of union catalogues, harvesting MPEG-7 data from multiple sources.

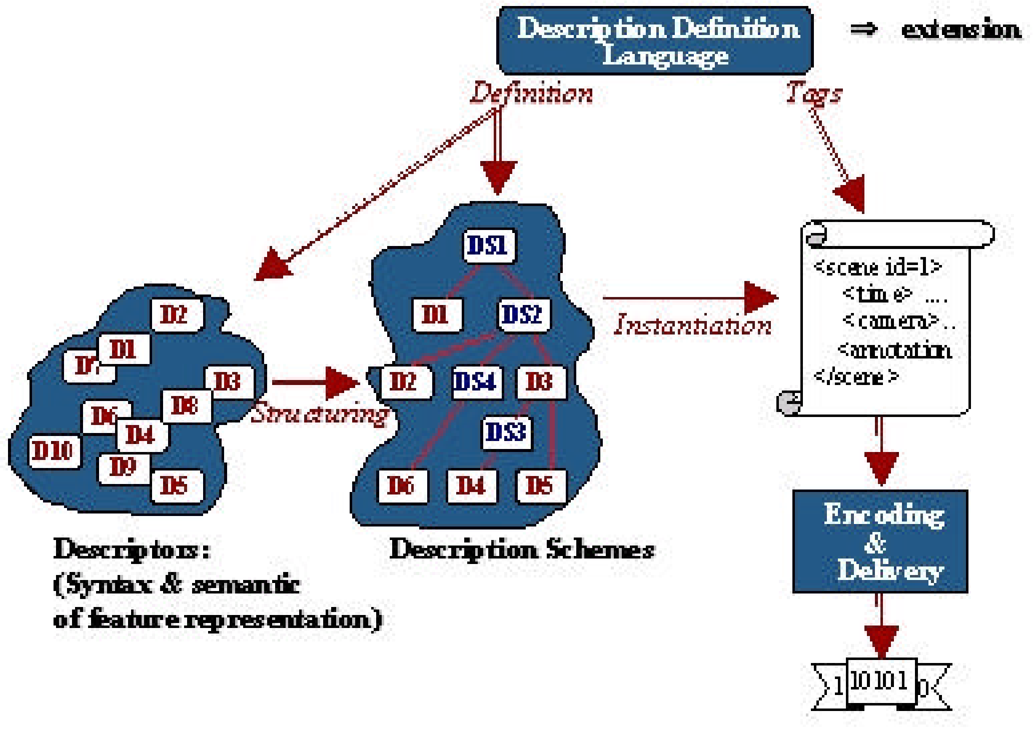
When the features are stable, it will be worth computing them. When archive users can find information in other archives because their metadata is in the MPEG-7 format, then there will be motivation for all of us to consider the same step.

MPEG-7 (Multimedia Content Description Interface)[28] provides a standardised content description for various types of audio/visual material (audio, speech, video, pictures…). It allows a quick localisation of the content. The standard is not restricted to the mentioned materials but can be also used to describe other aspects (e.g. user preferences…). Nevertheless only the content is described by this means, but neither description generation (e.g. indexing…) nor description consumption (e.g. search…). Examples of other application domains beside the a/v material related ones are:

- Tele-shopping.
- Intelligent multimedia presentations.
- Educational applications.
- Surveillance and remote sensing.
- Biomedical applications.

The so-called Data Definition Language (DDL) includes Descriptors (Ds) and Description Schemes (DSs), which allow the structuring of information related to content objects

Figure 2: MPEG-7 Description Definition Language[29]



.

3.7. Broadcast standards –EBU, SMPTE, AES, MPEG, MXF

3.7.1. EBU

The European Broadcast Union has many standards, and is an important source of technical support and guidance for all its members.

Particular standards are:

EBU P/FRA: The EBU panel on Future Radio archives produced three standards:

- The Capturing Report for Preservation Metadata:

EBU Technical Document: tech 3285_s2 Specification of the Broadcast Wave Format A format for audio data files in broadcasting: Supplement 2: Capturing Report

- A document giving general technical guidance for audio preservation transfers (so not a metadata document, but listed here for completeness):

EBU R105: Digitisation of programme material in Radio Archives

- A document on metadata, giving guidance to the use of Dublin Core in a radio archive context:

EBU Tech 3293: EBU Core Metadata Set for Radio Archives[30]

The use of core metadata is strongly advised. Whatever else an archive or preservation project must pay attention to, it must at least attend to core metadata. Using a standard for the core metadata can only help: it give clear definitions, there is a lot of guidance, and the resultant metadata will have maximum future usability. Core metadata is fully conceptually compatible with other EBU metadata work, and with SMPTE work. It just needs mapping, which is not a difficult issue (the metadata of three archives were mapped to a common standard as a small part of the work of the PRESTO project, as discussed above).

EBU P/AFT:

BWF: The Broadcast Wave Format TECH. 3285 (plus three supplements). This simple standard is now widely adopted, and provides a common file format for audio across broadcasting. It is equally useful for any audio archive using a file format. Without a common file format, file transfers are a muddle. Without the extra information in the BWF, broadcasters wouldn't be able to go 'straight to air', the file would be poorly identified, and there would still be a muddle. The BWF is a very simple standard, but all the more useful for its simplicity.

It holds not just one but two universal identifiers, the USID and the UMID (see Section3.9). Other information about BWF is on the EBU website: http://www.ebu.ch/pmc_bwf.html

FILM

Preservation and Reuse of Motion Picture Film Material for Television: Guidance for Broadcasters TECH. 3289-E March 01:

This very large and informative standard provides a wealth of information for the handling of film, its storage and its preservation requirements. Again, not a metadata issue, but vital for any film archive.

EBU P/META:

Another initiative was the PMC Project P/META (Metadata exchange standards). It was a project of the EBU (European Broadcast Union) led by BBC. The objective of that project was to standardise the structuring of media related information (either somehow associated with the media in separate data repositories or embedded in the media itself). The exchange of media items should benefit from that project. The work in this project can be seen as complementary to the various activities from EBU and SMPTE (e.g. Metadata Dictionary, UMID) as well. The main tasks of this project[31] are as follows:

- To establish understanding between EBU members of the media-related data interchange requirements of media commissioner/publishers (broadcasters), suppliers (producers) and consumers, using the BBC Standard Media Exchange Framework (SMEF) as the core information architecture.
- To validate and extend the SMEF model as appropriate against members’ requirements in terms of data and process, noting local synonyms (or translations), to create an “E-SMEF”. This would extend the thinking to the development of a commercial process framework for exchange of media between EBU members.
- Using E-SMEF, to apply emerging SMPTE metadata standards to the production and broadcast or distribution process, and study the feasibility of creating and adopting common exchange formats for essence and metadata.
- To establish understanding of the use of unique identifiers in metadata e.g. the SMPTE UMID, as a crucial linkage tool between unwrapped data (metadata) and wrapped or embedded metadata in media files or streams, and develop protocols for their management between members.

- As an aid to commercial and system interoperability between members, and in co-operation with standards bodies in related industries such as music and print publishing, to collate all relevant unique identifier schemes and map them against each other. This could be in collaboration with the EU INDECS project[32] and the DOI Foundation[33], and extend to cover their data models too.

3.7.2. SMPTE

The SMPTE Metadata Dictionary (SMPTE 335M-2000) is a reference book of audio-visual descriptors. These descriptors cover the entire production chain (pre-production, postproduction, acquisition, distribution, transmission, storage and archiving). A hierarchical registration of metadata items is possible through a general scheme. This scheme uses a universal label. Different description sets from other activities were combined into one common set.

The dictionary is made up of 10 categories dealing with the different aspects to be described. The number of categories can be increased to 255 if necessary. The data are encoded in the KLV (Key-Length-Value) protocol[34.] The SMPTE Universal Label is taken as the key. The automatically created length is according to ISO standards and the value is taken from the metadata dictionary.

The Unique Material Identifier (SMPTE 330M-2000) describes the format of a unique identifier for material like video, audio and data. The identifiers referring to that standard are created locally (thus not asking a general database for a registration) but still globally unique. This is a major difference to other identification methods. The reason why this uniqueness is possible lies in the fact that the identifier is made up of 2 parts: a Basic UMID and the Signature metadata. The Basic UMID contains the universal label, the length, the instance number and the material number. The Signature metadata is made up of time/date information, spatial coordinates, country and organisation codes and the name of the creator.

3.7.3. BBC SMEF

The SMEF™ (Standard Media Exchange Framework)[35] is a data model, which allows the description of all information related to the production, development, use and management of media assets. The model offers a semantic and a logical view on the items, logical clusters of items and the relationships in between the clusters. The model consists of two parts: a data dictionary defining the entities and attributes and a number of entity relationship diagrams, which show the structure in the form of relations between the entities and also the cardinalities in these relations.

It is a development of the BBC but has general validity and could therefore be used by other organisations dealing with media asset development and management issues as well. The model was not done as a standalone solution but also takes into account work done by other groups (e.g. EBU and SMPTE). Standards developed by those groups will be also incorporated in the future work on SMEF (or vice versa).

There is some lack of detail but considerable work is continuing on making SMEF and EBU P/META practical models for metadata processing. Future versions of the model will include a larger number of entities (covering the business area in a more accurate way). The attributes of various entities will be also increased. For each attribute more information will be added (e.g. format, validation rules, allowable value sets etc.). This will allow integration of more automated mechanisms.

3.7.4. EBU-SMPTE

As seen from above several activities exist in that area, and are running in parallel. In addition the fact that a lot of new distribution channels to customers are established (this is due to convergence of computer and communication technologies on one side and television on the other side) required the initiation of a common activity. The EBU and SMPTE formed a working group, the joint Task Force for the "Harmonisation of Standards for the Exchange of Television Programme Material as Bit Streams"[36.] Two major tasks were assigned to that Task Force: the production of User Requirements for implementing new technologies, which stay valid for at least one decade and fundamental decisions, which will in the end result in standards. These standards shall support future systems, which will be a consequence of the requirements.

Two reports were produced, the User Requirements (published in April 1997) and the Final Report (August 2000). The Final Report[37] resulted from a co-operation of more than 200 experts all over the world (Europe, North America, Australia and Japan). Within that report sections covered Systems in general, Issues of Compression, Transfer Protocols and of course Wrappers and Metadata. The report should be seen as a general guide for all following activities.

3.7.5. MXF

The BWF audio format has shown how vital and valuable it is to have an agreed file format. Progress from physical media and “archives with shelves”, to electronic media and archives with servers (and their tape-robot backup systems) requires agreement on simple, open, general file standards. MXF is the leading project to develop such a standard for video.

A considerable number of various multimedia file formats is in use during the production. These either correspond to standards, or are standards themselves, or are proprietary formats of specific systems. As a consequence transformations respectively conversions take place. A quality problem arises from that because one incorrect conversion may effect the things visible on the screen in a malicious way. For overcoming this problem the Advanced Authoring Format Association[38] develops a standard. This is done with the help of major industrial partners like Sony, Microsoft, BBC, CNN etc. The Advanced Authoring Format (AAF) is a software implementation of metadata and labels developed by SMPTE. The software is a wrapper application. It is available in the form of software development kit. The focus is on digital production but not on the delivery processes. Various types of essence can be handled through AAF (video, audio, MIDI, MPEG…). The mechanism used for describing the formats is called Media Exchange Format (MXF). This format is created in co-operation with the Pro-MPEG Forum[39].

MXF has been greatly enhanced by EC support (project G-FORS[40]). This project has advanced the standard very considerably, providing very full and clear documentation, and providing end-to-end test demonstrations of the technology. MXF is now (May 2002) undergoing the SMPTE balloting process.

3.8. Projects:

Metadata is an aspect of much EC-sponsored project work, and other national and internation activity. This paper is meant to be of practical value to audiovisual archivists, so it is not a survey of this entire field of activity. Two relevant projects will be mentioned

3.8.1. CEDARS

CEDARS[41] is about METADATA FOR DIGITAL PRESERVATION. It differs from PRESTO in that PRESTO concentrates on analogue preservation. Therefore CEDARS, and the OAIS initiative to which CEDARS is closely linked, is important because transfer from analogue to digital is not the whole preservation story – though it may well be the single most important step. CEDARS

3.8.2. PRESTO

The PRESTO[42] project surveyed use of metadata in 10 main European archives, and these results were reported in Deliverable 3.1. Some general findings were:

Some information items are used in all. The following list shows the common items:

- Tape number and bar code (e.g. = Material-Number), Archiving-Number, Production Number.
- Broadcast-Date / Channel.
- Title, Subtitle, Episode No.

- Duration / length.
- Recording format/standard.
- Colour, Ratio, Sampling, Sound.
- Quality of Picture/Sound.
- Usage of item
- Technical comments

The other main findings of PRESTO concerning actual and planned metadata usage are:

- The work related to the rights of materials is a major cost driver in many archives. Simplifications in this area can save quite some money in the future.
- Additional metadata should be collected in the future to provide additional savings.
  - At item level (transfer dates, technical notes and track listings).
  - New tape number or file number.
  - Corrected duration.
  - Correction of content if required.
  - Timecode and key frames.
  - Thumbnail pictures.
  - Genealogy of content.
  - Transcription quality control metadata.

- Legacy databases and local applications exist for various purposes but they do not offer the information in a standardised way. In case of more than one database in an archive (e.g. an additional rights database…) those database are not always linked and the exchange between those databases is often also somewhat difficult.
- The range of information provided with the material going out of the archives differs among the various archives from very limited sets to rich offerings, which can even be specified by the archives' customers. This even reaches a point where prints of catalogue-descriptions or copies of recording-forms can be requested. Making available such information with delivered material could be a benefit as well.
- Not all archives have so far made available the public access to their collections. Some of them even have a very restricting system (e.g. phone and fax for information about material and ordering…).
- Some of the archives within the PRESTO group are national ones. This means that they receive a lot of material from other institutions and individuals. The number of such items is at the moment greater than the archives can catalogue in their work and this is not going to decrease in the future. To get such material and the corresponding descriptions into the systems it will be necessary to use more standardised ways for cataloguing or at least for exchanging the metadata, which can then be easily fit into the existing storage systems.

There is other work in PRESTO, but the main metadata deliverable is this one, and I will resist the impulse to have this paper cite itself.

3.9. Identifiers

The principal universal identifiers coming up for standardisation within a broadcasting context are the USID and UMID, mentioned in Section 2.5.1.

Fuller information on identifiers and resource locators is on the IFLA website:

Identifiers: http://www.ifla.org/II/metadata.htm - identifiers

URLs and related concepts: http://www.ifla.org/II/metadata.htm - urls

3.10. Related areas of standardisation:

As companies and archives integrate by electronic technology, the role of metadata grows. The following areas are all probably as large in their own right as the area of standard archive metadata. This document cannot cover these areas, but will just give a paragraph of explanation and a key reference for each.

Consumer access: with digital broadcasting, satellite transmissions providing literally hundreds of channels, and technology emerging to greatly increase the automatic recording of material according to consumer choice, there is a real problem getting from the broadcaster to the consumer. At present there is minimum identification in the SI (Signal Identification) information, part of the DVB (Digital Video Broadcasting) standard. Richer information, amounting to an online programme guide or listings magazine, is provided by the EPG (Electronic Programme Guide). Unfortunately this is proprietary rather than standardised. However there is activity under the programme TV Anytime[43] to provide and increase standardisation in this vital area.

Digital Rights Management: In the BBC model:

essence + metadata = content

content + rights = an asset

There can be little electronic distribution of media without digital rights management. When this issue is studied, it soon turns into a metadata question: what categories of rights, what categories of usage, what structure (taxonomy) is needed to capture and manager the information? One project currently working on a rights taxonomy is Amicitia[44]

At present, DRM is dominated by the content providers, particularly in the cinema and music industries (because they will distribute material on DVD an CD which will be very susceptible to cloning and “perfect theft”, a much worse problem that for pirate VHS tapes). Broadcasting is not a dominant factor currently, and commercial aspects of archive management are a very small part of the picture.

E-commerce: Making a fortune from the web has gone through a boom and bust and perhaps is now again starting a boom cycle. The main point of significance for archivists is that the technology is being developed to sell simple commodities to a mass market, as in the Amazon book sale site[45]. Archives interested in this market need to find ways to use metadata (and related technology such as keyframes) to describe their offerings in a way that fits into the technology being developed for the mass market. If selling clips can be made to look like selling books, then there is a direct path from the archive to e-commerce – if the rights can be sorted out!

Networks: Finally, our greatest dependence in new technology is the network issue. If Henry Ford had shown people traffic jams before trying to sell cars, we might never have had an automotive industry.

The guidance from the EBU is clear: media needs to move from synchronous lines to IP (Internet Protocol) data lines, which are far cheaper. However there is huge internal competition within broadcasting for who gets the ‘fast lane’ on the data network. At present the BBC and most other broadcasters are in the very poor position of having the data network staffed and planned solely by persons who are concerned with ‘business data’, which to the rest of us means email. The guidance from the EBU[46] is very clear: media needs a separate data network. This is not a metadata issue, except that it is another example of the importance of separating data (audio and video) from metadata (email).

4. Conclusions

It is the authors hope that the information in this document and its partner, D7.1 Metadata Reference Process, will take the pain out of metadata, and encourage those involved in preservation work to make metadata work for them rather than being yet another problem.

The intent of these documents is to show that metadata saves money and raises quality in preservation work, and that an investment in automation of metadata is the first and most important stage of process transfer automation.

Finally, an investment in metadata itself, making core archive metadata electronically available in a standard form, opens the doors to full exploitation of new technology, for direct electronic access: the archive on the desktop.

References

General Metadata Overviews:

1- Annemieke de Jong: Metadata in the audio-visual production environment; ©2000 Netherlands Audiovisueel Archief

2- Jean-Noel Gouyet: Outils de Digital Media Asset Management A 350-page INA Recherche & Experimentation Report - January 2001.[47]

3- Henny Bekker, Ivana Belgers, Peter Valkenburg; *Inventory of Metadata for* Multimedia; September 2000[48]

4- The Getty Research Institute: Introduction to Metadata[49]

5- On the official National Library of Australia site: Meta Matters: What is Metadata?[50]

6- General information on standards: http://www.library.ucsb.edu/subj/standard.html

Quick guide to reference material:

AAF (multimedia authoring file format)	http://www.aafassociation.org
Broadcast Wave Format (EBU audio file format BWF)	http://www.ebu.ch/pmc_bwf.html
Dublin Core (General basic metadata)	http://uk.dublincore.org/
EBU general	http://www.ebu.ch/technical.html
EBU P/AFT (BWF, etc)	http://www.ebu.ch/pmc_aft.html
EBU P/META (metadata)	http://www.ebu.ch/pmc_meta.html
EBU P/FRA (radio archives)	http://www.ebu.ch/pmc_fra.html
FIAT (International Federation of Television Archives)	http://fiatifta.org/
IASA (International Association of Sound and Audiovisual Archives)	http://www.llgc.org.uk/iasa
MPEG-7 (metadata standard)	http://www.cselt.it/MPEG/standards.htm
MXF (Project G-FORS) SMPTE video file format	http://www.g-fors.com/pub/rootframeset.htm
PRESTO public papers	http://presto.joanneum.ac.at/index.asp
Pro-MPEG Forum	http://www.pro-MPEG.org
SMEF (BBC metadata)	http://www.bbc.co.uk/guidelines/smef
SMPTE Dictionary	http://www.smpte-ra.org/mdd/index.html
XML	http://www.xml.com/pub/a/98/10/guide1.html

[1] http://www.imsproject.org/metadata/mdbest01.html#_Toc459646401

[2] http://www.w3.org/Consortium/Legal/ipr-notice-20000612#Legal_Disclaimer

[3] http://foldoc.doc.ic.ac.uk/foldoc/index.html

[4] http://www.ebu.ch/tech_text_r99-1999.pdf

[5] SMPTE UMID: http://www.smpte.org/smpte_store/standards/index.cfm?scope =0&CurrentPage=15&stdtype=smpte

[6] http://www.bic.org.uk/uniquid

[7] All public PRESTO papers are at http://presto.joanneum.ac.at/index.asp

[8] S.-F. Chang, W. Chen, and H. Sundaram, Semantic Visual Templates: Linking Visual Features to Semantics, ICIP'98, Workshop on Content Based Video Search and Retrieval, Chicago IL, Oct 10-12 1998

R. Brunelli, O. Mich, C. M. Modena, A Survey on Video Indexing, J. of Visual Communication and Image Representation 10, pp.78-112, 1999

Shi-Fu Chang and H. Sundaram, Structural and Semantic Analysis of Video, Proceedings of ICME 2000: IEEE International Conference on Multimedia and Expo New York, USA, July 30 - August 2, 2000

H. Wactlar, Informedia - Search and Summarization in the Video Medium, Proceedings of Imagina 2000 Conference, Monaco, January 31 February 2, 2000

[9] Thisl: http://www.dcs.shef.ac.uk/research/groups/spandh/projects/thisl/

[10] Olive: http://twentyone.tpd.tno.nl/olive/

[11] http://www.getty.edu/research/institute/standards/intrometadata/1_introduction/index.html

[12] http://www.nla.gov.au/meta/intro.html

[13] http://www.getty.edu/research/institute/standards/intrometadata/2_articles/index.html

[14] Preservation Of The Bbc Radio 1 Archive; Presentation To Aes 20th International Conference. Author: Allan King, BBC London; allan.king@bbc.co.uk

[15] EBU Technical Document: tech 3285_s2 Specification of the Broadcast Wave Format
A format for audio data files in broadcasting: Supplement 2: Capturing Report
http://www.ebu.ch/tech_t3285_s2.html

[16] Library of Congress, 1994, USMARC format for bibliographic data including guidelines for content negotiation, Network Development and MARC standards office, Library of Congress: Washington DC

[17] http://www.niso.org/standards/resources/Z3950_Resources.html#info

[18] http://lcwb.loc.gov/z3950/agency

[19] Paul Miller: Metadata for the masses: describes Dublin Core and means by which it can be implemented; http://www.ariadne.ac.uk/issue5/metadata-masses/

[20] EBU R105: Digitisation of programme material in Radio Archives
http://www.ebu.ch/tech_t3293.html

[21] Open Archival Information System http://www.rlg.org/longterm/oais.html

[22] Dublin core for museum informatics; various papers: http://www.cimi.org/publications.html

[23] CEDARS Digital Archive project: Metadata:
http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html#div1

[24] Reference Model for an Open Archival Information System (OAIS);
http://www.ccsds.org/RP9905/RP9905.html

[25] IASA Technical guideline TC-03: http://www.llgc.org.uk/iasa/iasa0013.htm

[26] IASA Cataloguing Rules: http://www.llgc.org.uk/iasa/icat/index.htm

[27] XML http://www.xml.com/pub/a/98/10/guide1.html

[28] MPEG-7 main page; GMD - Forschungszentrum Informationstechnik GmbH:

http://www.darmstadt.gmd.de/mobile/MPEG-7/index.html

[29] Philippe Salembier: Status of MPEG-7: the Content Description Standard; International Broadcasting Conference Amsterdam, The Netherlands, September 8, 2000; Universitat Politecnica de Catalunya, Barcelone

[30] EBU Tech 3293: EBU Core Metadata Set for Radio Archives

http://www.ebu.ch/tech_t3293.html

[31] European Broadcasting Union: PMC Project P/META (Metadata exchange standards):

http://www.ebu.ch/pmc_meta.html

[32] http://www.indecs.org

[33] http://www.doi.org

[34] http://www.smpte.org/stds/s336m.pdf

[35] SMEF™ DATA MODEL v1.5; ”2000: British Broadcasting Corporation.

[36] Joint EBU / SMPTE Task Force on User Requirements for the Exchange of Television Programme Material as Bit Streams

[37] European Broadcasting Union: Task Force for Harmonized Standards for the Exchange of Program Material as Bitstreams; Final Report: Analysis and Results; July 1998:

http://www.smpte.org/engr/tfrpt2w6.pdf

[38] http://www.aafassociation.org

[39] http://www.pro-mpeg.org

[40] MXF (Project G-FORS) SMPTE video file format;

http://www.g-fors.com/pub/rootframeset.htm

[41] http://www.leeds.ac.uk/cedars/

[42] http://presto.joanneum.ac.at/index.asp

[43] http://www.tv-anytime.org/

[44] http://www.amicitia-project.de/

[45] http://www.amazon.com/

[46] Maurizio Ardito; Presentation at EBU Technical Assembly, Larnaca, Cyprus, April 2002

[47] http://www.ina.fr/Recherche/Etudes/evaluations/Etude5/Sommaire.fr.html

[48] http://www.surfnet.nl/innovatie/surfworks/doc/mmmetadata/

[49] http://www.getty.edu/research/institute/standards/intrometadata/1_introduction/index.html

[50] http://www.nla.gov.au/meta/intro.html