Keeping Emulation Environments Portable

FP7-ICT-231954

Preliminary document analyzing and summarizing metadata standards and issues across Europe

Deliverable number	D3.1
Nature	Report
Dissemination level	PU
Delivery Date	**Due**: 31st July 2009
	Actual: 30th July 2009
Status	Final
Workpackage number	WP3
Lead beneficiary	UPHEC
Author(s)	David Anderson (UPHEC) Janet Delve (UPHEC) Dan Pinchbeck (UPHEC) Getaneh Agegn Alemu (UPHEC)

Document history

Revisions

Version	Date	Author	Changes
1.1	28th July 2009	David Anderson	Response to Reviewer (Steinke)
1.2	30th July 2009	David Anderson	Response to Reviewer (Sharpe)

Reviews

Date	Name	Result
24th July 2009	Tobias Steinke (Deutsche Nationalbibliothek)	Approved (minor modifications)
30th July 2009	Robert Sharpe (Tessella)	Approved (minor modifications)

Signature/Approval

Approved by (signature)	Date

Accepted by at European Commission (signature)	Date

Executive Summary

This document is a report on the state-of-the-art in metadata standards and approaches in Europe. Metadata are widely recognized as a critical component of digital preservation and it is typically the case that within individual cultural heritage organizations numerous different metadata schemes are employed, each of which aims to capture particular aspects of digital objects. KEEP is particularly focused on emulation as a digital preservation strategy and addresses directly dynamic digital objects. Emulation places unique demands on metadata. In addition to holding fairly general information about preserved digital objects (format etc.). It is often possible to determine this kind of information by closely examining stored digital objects. However an emulation approach also requires us to have at our disposal a great deal more detailed information on environments (e.g, creating application, operating system etc.). Information of this sort can to some extent be derived – for example, if a digital object is determined to be an AppleScript application we may reasonably infer that it was produced on an Apple platform rather than on an IBM PC clone. However, such inference is often incomplete and frequently ambiguous leaving us in the position that we cannot say exactly which application created the digital object or know for sure the target platform(s) for which it was originally intended. In order to know which emulator is appropriate or best to run a given digital object, it is precisely this sort of information which is needed

Within that context we need to investigate whether there is a need to develop new preservation approaches to record metadata pertinent to emulation. This document discusses the various digital preservation strategies currently employed, and assesses the extent to which they address the demands imposed by dynamic objects. We examined the role played by 'environment' or 'technical' metadata in current metadata standards and practice in three national libraries and a computer games museum: Bibliothèque nationale de France (France), Deutsche Nationalbibliothek (Germany), Koninklijke Bibliotheek (Netherlands) and Computerspielemuseum Berlin (Germany). Finally we raise a number of issues that need to be addressed by KEEP (and its successors) in the future.

The literature on preservation metadata standards shows that there has been very little effort expended directly in the development and implementation of preservation metadata to support digital preservation strategies based around emulation. Unsurprisingly, given that the main goal of libraries is to provide access to their digital collections with efficient search systems, their primary interest has been in the development of descriptive metadata such as author, title, subject, publication and date.

The most cited preservation metadata standard is PREMIS which is a result of years of work of international experts under the OCLC/ RLG working group. The PREMIS data dictionary is a high level definition of metadata schema for preservation purposes. It defines core implementable metadata which should mean that the PREMIS metadata dictionary is not tied to any specific preservation strategy but we found that in practice it supports migration more easily than emulation.

It is essential that emulation-based digital preservation strategies develop scalable, interoperable metadata schema which capture enough detail to record core information about objects, and their hardware and software environments. Emulation metadata must also record information about rights, provenance, and authenticity.

In the development of a metadata schema for emulation-based digital preservation strategies, the OAIS conceptual model should be able to serve as a reference model to assist scalability and interoperability.

Grid computing is currently a favoured approach for web archiving. The ramifications of this for emulation should be borne in mind;

Emulation is in its infancy in terms of use by major library / archival institutions; however these bodies are clearly stating an urgent need for this preservation strategy to deal with burgeoning collections of evermore complex and dynamic digital objects;

OAIS, METS and PREMIS are standards around which the three national libraries can coalesce, even though each is likely to have their own instantiation.

File format recognition software such as PRONOM could play a vital part in any future emulation system by automatically providing technical metadata for a good proportion of complex digital objects, and this could help in the uptake of emulation by libraries who might otherwise find it not sufficiently automated.

The games preservation community has thrown up some interesting work. Huth's model represents the only dedicated and systematic model for game preservation metadata currently available, and further study should be undertaken to properly analyse compatibility with the PREMIS extension being considered as the core metadata structure of KEEP. It is also the only model that aims to specifically include emulation and detailed run-functionality technical data. Concerns are the complexity of the model and the impact of this upon a non-automated ingest procedure. Equally, Huth's model does raise very starkly the sizeable issue with cross-dependencies and object- extensions/alterations (in the form of patches, commercial extension packs, cracks and mods) that are so common in this medium, and of real importance for the preservationist.

Commercial game sites, typified by Gamespot, appear to offer accurate but limited metadata about recent releases and may provide information about new objects being ingested. It may be worth considering farming this data now whilst it is readily available, for later archiving of the objects. In other words, although most of the objects detailed by these sites are currently too complex for robust emulation (although the emulation community have produced emulators for many current platforms, mainly consoles, with all the legal issues surrounding these we might expect), there is no reason why objects could not be ingested for later emulation, in which case, the descriptive data supplied in these sites would become useful.

MobyGames is far more practically useful at this stage as, although unsystematic and community-driven, it does engage with older games that are more likely to be suitable for emulation via the KEEP framework. Again, technical metadata is extremely limited, but the wealth of descriptive metadata available does suggest this is a resource not to be overlooked. In particular, supplementary descriptive metadata such as developer credits on the site extend Huth's model

Abandonware sites, although often holding and distributing material of a somewhat fuzzy legal status, and having limited and highly unsystematic metadata structures, nevertheless may offer access to both objects and simple emulation metadata that could be of use to KEEP. The fact that Abandonia, for example, explicitly suggests suitable emulators for objects in its archive should be noted. Further, regardless of the problems in terms of limited metadata and legality, it is community-driven sites like these that have probably done more, in international terms, to preserve computer games than any other preservationists, including national libraries and archives.

Even if KEEP requires a greater robustness to its metadata and preservation strategies, we should actively seek dialogue and aim to supplement and enhance the large, if shallow, body of information available through these initiatives.

Over that last decade there has been considerable effort expended on defining preservation metadata elements, the overwhelming majority of which have been intended to support migration strategies. While a few attempts have been made to define environment metadata they have insufficient specificity and detail to be used for the emulation framework at the heart of the KEEP project.

Table of Contents

Section	Title	Page
1	**Introduction**	10
2	**Data Preservation Strategies**	12
2.1	Bitstream Copying (Replication)	12
2.2	The Paradox of Migration	13
2.2.1	Hardware Migration (Refreshing)	13
2.2.2	Software (File Format) Migration	14
2.2.2.1	Format Simplification	14
2.2.2.2	Migration on Request	14
2.3	Emulation	14
2.4	KEEP as a Hybrid Approach	15
3	**Preservation Metadata Initiatives and Standards**	16
3.1	Metadata Registries	16
3.2	The many roles metadata play for digital preservation	16
3.2.1	Metadata Types	17
3.2.2	Preservation Metadata Roles / Approaches	17
3.3	Open Archival Information System (OAIS)	18
3.4	Preservation Metadata Projects and Initiatives	21
3.5	The Research Libraries Group (RLG)	21
3.6	PANDORA Logical Data Model (National Library of Australia)	22
3.7	Preservation Metadata Standards Framework (National Library of New Zealand)	24
3.8	Networked European Deposit Library (NEDLIB) Metadata Elements	27
3.9	CEDARS Preservation Metadata Elements	30
3.10	PREMIS (PReservation Metadata Implementation Strategies)	35
3.11	PREMIS implementation	37
3.12	The need to extend PREMIS	39
3.13	LMER (Long-term preservation Metadata for Electronic Resources)	42
3.14	METS (Metadata Encoding and Transmission Standard)	43
4	**Relevant Framework 7 Projects**	46
4.1	LiWA - Living Web Archives.	46
4.2	PARSE.Insight (Permanent Access to the Records of Science in Europe)	46
4.3	PROTAGE (Preservation Organizations Using Tools in Agent Environments)	46
4.4	PrestoPRIME	47
4.5	SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg)	47
5	**Relevant Framework 6 Projects**	48
5.1	CASPAR (Cultural, Artistic and Scientific Knowledge Preservation, for Access and Retrieval)	48

5.2	Digital Preservation Europe	48
5.3	P2P-FUSION - Peer to Peer Fusion.	48
5.4	PLANETS - Preservation and Long-term Access to our Cultural and Scientific Heritage.	48
5.5	PRESTOSPACE - Preservation towards storage and access.	49
6	**Other Relevant Projects and Developments**	50
6.1	Europeana	50
6.2	DP Fuse (Digital Preservation Functions and Services)	
6.3	The European Library (TEL)	50
6.4	nestor (Network of Expertise in Long-Term Storage of Digital Resources)	50
6.5	The PRONOM Technical Registry	51
6.6	The Global Digital Format Registry (GDFR)	52
6.7	The Unified Digital Format Registry (UDFR)	52
6.8	IBM Preservation Manager	53
7	**Institutional Issues**	54
7.1	Brief Institutional Overview	54
7.2	Current European Research Projects / Collaboration	54
7.3	Collection Policies	54
7.4	Current / Future Preservation Systems at the BnF	55
7.5	Current / Future Preservation Systems at the KB	56
7.6	Current / Future Preservation Systems at the DNB	57
7.7	Current / Future Preservation Systems at the CSM	60
8	**Preservation Metadata For Games**	61
8.1	Recent research in Germany	61
9	**Non-Preservation Metadata For Games**	63
9.1	MobyGames	61
9.2	Gamespot	66
10	**Abandonware Sites**	67
10.2	Abandonia	67
10.3	Home of the Underdogs (HoTU)	68
11	**Abbreviations**	69
12	**Bibliography**	74
13	**Endnotes**	79

Tables included in this report

Title	Page
Table 1: Information Produced Worldwide	11

Figures included in this report

Title	Page
Fig.1. The OAIS Functional Model	19
Fig.2. The OAIS Information Model	20
Fig.3. RLG metadata elements	22
Fig.4. PANDORA Data Model	23
Fig.5. NLNZ metadata schema (partial) for the 'Object' metadata element	26
Fig.6. NEDLIB Metadata Elements (Representation Information)	29
Fig.7. NEDLIB Metadata Elements (Preservation and Description Information)	29
Fig. 8. CEDARS Metadata Elements	33
Fig. 9. CEDARS metadata elements	34
Fig. 10. Details held by PRONOM on the Apple Macintosh Microsoft Word 6.0 file format.	37
Fig. 11. PREMIS Data Dictionary Source	41
Fig. 12. LMER metadata elements	43
Fig. 13. METS Sections	44
Fig. 14. Example View Path	53
Fig. 15 A dimension of the new Reference Architecture for SHAMAN	59

1 Introduction

The purpose of the KEEP project is to help facilitate universal access to our cultural heritage by developing flexible tools for accessing and storing a wide range of digital objects – primarily in the form of files created on computers.

Digital objects present the archivist or librarian interested in preservation with a variety of problems not all of which are entirely unfamiliar, such as the responsibility to record provenance information (Lundy, 2008) or to provide safe handling and storage conditions (Marcum & Friedlander, 2003). However digital objects are quite different from the sort of items stored in libraries or museums before the 20th century. Being immaterial they give rise to questions which simply do not apply to books, or sheet music and which are played out, for example, in discussions about what counts as a 'document' in the digital context (Buckland, 1997) or what constitutes authenticity in a digital environment (A. Smith, 2000b), (Geser et al., 2002),(Diessen & Werf, 2002). Abby Smith noted in 2000 a blurring of the distinction at the digital level of objects which are quite distinct in the physical world. As she puts it:

> "In the end, museum objects and library items are indistinguishable from one another when transferred to digital form. A digital Blue Boy and a digital Huckleberry Finn share the same behaviors, demand the same creation of metadata and management tools, reside on the same network, and are retrieved onto the same computer screen."(A. Smith, 2000a, p. 34).

The purpose of preservation is to ensure protection of information of enduring value for access by present and future generations (Conway, 1990, p. 206). There is some disagreement about how long digital objects need to be preserved in order to count as having been preserved in the "long term" with, for example (Exon, 1995) arguing not unreasonably that nothing less than several hundred years would suffice, (Janée, Mathena, & Frew, 2008) suggesting one hundred years while at the other end of the spectrum (Verheul, 2006, p. 20) opines that "Long-term means five years or more". What is not in doubt is that digital objects are relative newcomers on the information landscape. The development of the digital computer in Manchester (UK) in 1948 (Anderson, 2007) marks the earliest reasonable starting point for the generation of the class of digital objects which is most directly of interest to KEEP – computer files. In a little under seventy years the volume and variety of digital objects which have sprung from this source has grown exponentially to the point where at the start of the millennium it was noted that "almost 800 MB of recorded information is produced per person each year." (Lyman & Varian, 2003).

Writing in 2005, Mackenzie Smith put this into context:

> "It took two centuries to fill the U.S. Library of Congress in Washington, D.C., with more than 29 million books and periodicals, 2.7 million recordings, 12 million photographs, 4.8 million maps, and 57 million manuscripts. Today it takes about 15 minutes for the world to churn out an equivalent amount of new digital information. It does so about 100 times every day, for a grand total of five exabytes annually. That's an amount equal to all the words ever spoken by humans." (M. Smith, 2005)

Printed media represents only 0.01 percent of the total which even allowing for some measure of disagreement about the precise significance of what (Lyman & Varian, 2000) have called "the paucity of print" (See Collins & Murroni, 1996 for an

alternative view) leaves no room for doubt as to the importance that digital information has come to have in the information universe.

Worldwide production of original information, if stored digitally, in terabytes circa 2002. Upper estimates assume information is digitally scanned, lower estimates assume digital content has been compressed.					
Storage Medium	2002 Terabyes Upper Estimate	2002 Terabyes Lower Estimate	1999-2000 Upper Estimate	1999-2000 Upper Estimate	% Change Upper Estimates
Paper	1,632	327	1,200	240	36%
Film	420,254	7,669	431,690	58,209	-3%
Magnetic	4,999,230	3,416,230	2,779,760	2,073,760	80%
Optical	103	51	81	29	28%
Total	5,421,221	3,416,281	3,212,731	2,132,238	69%

Table 1: Information Produced Worldwide Source (Executive Summary Lyman & Varian, 2003, p. 4)

The speed of the shift towards digital objects and away from physical objects would have been enough to present serious challenges to librarians and archivists even had it not been accompanied by an equally rapid evolution in hardware and software platforms. Computers are often characterised in 'generational' terms with crude lines of demarcation being drawn at those points where a fundamental shift occurred in the underlying technology for example from reliance on integrated circuits to the use of microprocessors. However this simplistic characterisation both understates the degree to which hardware platforms evolve within a single base technology and fails completely to represent the continuous nature of the evolution of software, computer applications, file formats and operating systems. A change in any of the hardware or software parameters has the capacity to undermine seriously an otherwise well-designed preservation strategy by rendering files created on one computer system completely unreadable on another.

2 Data Preservation Strategies

The speed of technological and organizational change in the digital environment represents the main stumbling block for the development of robust long-term digital preservation strategies and has led to a situation where we have been able to preserve written material over millennia but struggle to preserve digital information a few decades ago. The situation is nowhere more critical than with born-digital information because there is no analogue counterpart to fall back on if digital preservation fails.

A digital preservation strategy, according to the Online Computer Library Center, "details the types of activities that will be undertaken to ensure reliable preservation of digital content objects. These activities include:

- Assessing the risks for loss of content posed by technology variables such as commonly used proprietary file formats and software applications.
- Evaluating the digital content objects to determine what type and degree of format conversion or other preservation actions should be applied.
- Determining the appropriate metadata needed for each object type and how it is associated with the objects.
- Providing access to the content. There are several additional strategies that individuals and organizations may use to actively combat the loss of digital information." (OCLC, 2006, p. 5)

The detailed preservation actions needed for digital material vary significantly with the precise nature the type of digital object being preserved and are greatly dependent on the digital preservation strategy adopted (McLeod, Wheatley, & Ayris, 2006). There are various options available.

2.1 Bitstream Copying (Replication)

This refers to making an exact "bit-for-bit" copy of a digital object and may reasonably be thought of a necessary component of all digital preservation approaches. In the absence of complementary techniques (such as ensuring that replicated data is stored in multiple locations) bitstream copying cannot be considered as a long-term preservation technique as all the copies may be subject to the same physical threats such as software or hardware failure, intentional or accidental alteration, and environmental catastrophes like fire, flooding, etc. The notion that making multiple copies of digital files offers a measure of security underpins the LOCKSS (Lots of Copies Keep Stuff Safe) initiative based at Stanford University Libraries. LOCKSS is "an international community initiative that provides libraries with digital preservation tools and support so that they can easily and inexpensively collect and preserve their own copies of authorized e-content"[1]. It is a digital preservation Internet appliance in which preserving material in the collection is intertwined with the provision of access to the end user (Reich & Rosenthal, 2001). The LOCKSS initiative had led to a joint venture called CLOCKSS (Controlled LOCKSS) "with the world's leading scholarly publishers and research libraries whose mission is to build a sustainable, geographically distributed dark archive with which to ensure the long-term survival of Web-based scholarly publications for the benefit of the greater global research community."(CLOCKSS, 2008)

2.2 The Paradox of Migration

It has been observed that technological obsolescence represents a far greater threat to preserved digital objects than any inherent physical fragility of storage media (Gavrel, 1986; Mallinson, 1986; Preserving Digital Information, 1996). Migration appears to offer the opportunity to overcome technological obsolescence by making digital objects accessible on each successive generation of hardware but only at the expense of changing the bitstream which digital preservation strategies were devised to preserve unaltered. (Chen, 2001) Indeed over the long term, the bitstream of a migrated digital object may come to have no overlap whatsoever with the bitstream that was created during ingest. For example a program originally written in FORTRAN for an IBM Machine in the 1950s may have been migrated a number of times so that, in its present incarnation, is has been rewritten in QCL^2 in order to run on a quantum computer. The chances of significant overlap between the original bitstream and the migrated version are slim and get slimmer with each migration intervention.

It is beyond question that at some point in the future it may no longer be possible to migrate certain digital objects while retaining those features held to be of central importance. This state of affairs is far from optimal and has led a number of researchers to question whether another approach might be tried. KEEP is a practical step in an alternative direction.

A great deal of effort is inevitably expended by those responsible for preservation management to establish the core functionality and key features which need to be preserved across migration interventions. However, the choices that are made about which features to preserve depend to a very great extent on the particular preservation community doing the choosing. There are many different ways of viewing objects (digital or physical) and strategies that appear obvious from one perspective are completely unwarranted from another. As Winget put it:

> "If a writer chose to write hypertext fiction in hypercard, and it eventually got migrated to Word XP or Flash, is the work being honestly represented? To take a more strident example: would the Sistine ceiling convey the same meaning if the fresco layer were separated from the wall, transferred to canvas or wood, and hung in a museum? The technology is certainly there, but it would be a fundamentally different experience. Again, it becomes a question of whether the conservation/preservation community is trying to preserve access to the physical content of a work, or if they're trying to preserve access to its deeper meaning. It becomes a very sticky business wherein the conservator is making major artistic choices traditionally left to the artist." (Winget, 2005)

Migration is a model-orientated approach to preservation in that it compels us to stipulate on behalf of future generations, whose concerns and interests may inevitably be very different to our own, not only which digital objects should be retained but how we should view them.

2.2.1 Hardware Migration (Refreshing)

This is the process of the faithful duplication of a bitstream stored on one long-term medium (e.g., CD-RW) on another (e.g., DVD-RW) (Bearman, 1989). Refreshing is simply an extension to the process of bitstream copying. It is relatively limited as an effective preservation technique. As the Commission on Preservation and Access

and Research Libraries Group put it in 1996: “Refreshing digital information by copying will work as an effective preservation technique only as long as the information is encoded in a format that is independent of the particular hardware and software needed to use it and as long as there exists software to manipulate the format in current use. Otherwise, copying depends either on the compatibility of present and past versions of software and generations of hardware or the ability of competing hardware and software product lines to interoperate. In respect of these factors - backward compatibility and interoperability -- the rate of technological change exacts a serious toll on efforts to ensure the longevity of digital information.” (Preserving Digital Information, 1996)

2.2.2 Software (File Format) Migration

Migration, as the Task Force on the Archiving of Digital Information put it in 1996, “is the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology.”(Preserving Digital Information, 1996)

Migration may, in practice, amount to little more than refreshing but will often involve making alterations to the original bitstream in order make it compatible with a new generation of technology. (Waters & Garrett, 1996)

2.2.2.1 Format Simplification

Digital preservation would clearly be a great deal easier if the number of file formats being managed could be reduced and/or if the inherent complexity of the file structures involved could be simplified[3]. In some cases archives place restrictions the formats that they will accept for preservation, or convert the files presented to them into formats that they can maintain.

An alternative (but related) approach is the accept a wide range of formats but to ‘normalize’ them into a standard file structure which is expected to remain recognizable by computers long into the future. XML has shown considerable promise as a format by which this might be achieved.

2.2.2.2 Migration on Request

Migration on request was conceived by the CEDARS project. Original digital objects are preserved together with a migration tool that runs on a current computing platform. Users employ the migration tool to convert the original bitstream of their target digital object into a current format. Over time, the hardware platform on which the migration tool operates will become obsolete and the migration tool must either be replaced or migrated. In practical terms there is little difference between this approach and emulation. (Holdsworth & Wheatley, 2001) (Wheatley, 2001)

2.3 Emulation

Writing in 1999 Jeff Rothenberg said “There is as yet no viable long-term strategy to ensure that digital information will be readable in the future. Digital documents are vulnerable to loss via the decay and obsolescence of the media on which they are

stored, and they become inaccessible and unreadable when the software needed to interpret them, or the hardware on which that software runs, becomes obsolete and is lost."(Rothenberg, 1999)

His proposed solution was "to emulate obsolete systems on future, unknown systems, so that a digital document's original software can be run in the future despite being obsolete."(Rothenberg,1999). The feasibility of emulation as a complete preservation solution has been debated in the academic community (Granger, 2000). David Bearman has been particularly scathing on emulation as a 'magic bullet solution' which he believes fails to take account of significant bodies of prior literature, fails to "address the problems of maintaining electronic records, won't work as a strategy, and may encourage potentially dangerous wishful thinking."(Bearman, 1999)

2.4 KEEP as a Hybrid Approach

Certainly emulation represents a more 'source-oriented' approach to preservation than is afforded by migration. Assuming that we had available to us a suitable emulator for the original hardware platform for a given digital object, it would be possible to avoid having to make any of the difficult decisions concerning key features which are forced on the preservation community by the migration approach. As far as preserved digital objects are concerned, the job of preservation would involve nothing more than bitstream copying and refreshing. It is not possible to avoid entirely the need for migration but the focus of attention would, so to speak, turn to migrating emulators to new hardware platforms when the old ones became obsolete rather migrating the digital objects which we seek to preserve. This is a tempting strategy in that it requires that a few (relatively) emulators have to be migrated rather than many (other) digital objects. However emulators are complex programs and require detailed understanding of the hardware platform being emulated in order to ensure fidelity with the original hardware (and therefore software) performance. It is by no means certain that completely successful emulators will ever be produced for any or all hardware platforms. Certainly at the current stage of development there is no question of there being a single emulator which can support, for example, every program ever written for an IBM PC produced in 1982. For a given digital object it is usually a case of deciding subjectively which of a number of available emulators renders the most satisfactorily authentic experience. Emulators are programs that are written to run on specific hardware platforms and thus even were a 'perfect' emulator to be developed; the time will come when the hardware on which it runs becomes as obsolete as the hardware that it is emulating. At that point the choice is either to recursively run emulator A on emulator B or to migrate emulator A onto the next hardware platform.

Each of these approaches is problematic. KEEP is an attempt to make the process of moving emulators onto new hardware platforms less difficult by providing an emulator framework written in a highly portable computer language; a language so simple to migrate that few or no difficult choices about 'key features' have to be made again.

3 Preservation Metadata Initiatives and Standards

Metadata can be plausibly defined as any “structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage any other resource.” (NISO, 2004) Regrettably, the over use of the term without further clarification in a wide range of different contexts effectively robs the word of meaning. Consequently, it is quite usual to refine the use of the term by the addition of a disambiguating adjective.

In what ever way the term is construed, metadata is a vital component of any digital preservation strategy (Baker et al., 2003; Blanchi & Petrone, 2001; Brown, 2008; Heery & Wagner, 2002) so it is not surprising that there has been no shortage of metadata initiatives and projects aimed at developing preservation metadata schemas and standards. Indeed, without appropriate metadata any attempt to ensure the longevity and authenticity of digital objects cannot succeed.

Overall, little has been done about preservation metadata and most of the work that has been carried out was developed to support migration. Our review shows that the current preservation metadata schemas are not specific enough to be used for digital preservation strategies based on emulation. The focus has been overwhelmingly on providing descriptive and technical information about the digital object whereas the hardware and software environments on which the object was created and/or rendered have been substantially ignored. For digital preservation strategies where emulation is a key component, the hardware and software context is absolutely vital.

This report focuses on the Open Archival Information System (OAIS) Reference Model (ISO 14721:2003), the PREMIS Data Dictionary (version 1.0), Cedars (CURL Exemplars in Digital Archives), and the Metadata Encoding and Transmission Standard (METS). Some earlier attempts at preservation metadata development are also reviewed: the National Library of Australia’s PANDORA logical data model, the National Library of New Zealand’s preservation metadata standards framework, CEDARS (CURL exemplars in digital archives) and NEDLIB (Networked European Deposit Library).

3.1 Metadata Registries

One of the challenges faced by the preservation community is to continue to work towards greater convergence and interoperability. Some of the impetus for convergence within the cultural heritage domain has come from an increasing realisation that, especially when it comes to metadata, there is a considerable common involvement in content creation and networked service delivery. There is also a widespread desire to reduce or avoid completely any duplication of effort which has given support to the development of metadata registries (NISO, 2004); central locations where metadata definitions are stored and maintained.

A number of metadata registries exist which are particularly relevant for an emulation-centric digital preservation strategy. (See §6.5, 6.6, 6.7 & 6.8 below)

3.2 The many roles metadata play for digital preservation

There are many descriptors currently in use for metadata, covering their types and roles.

3.2.1 Metadata Types

Gartner draws attention to the "complex metadata requirements of digital objects, which include descriptive, administrative and structural metadata." (Gartner, 2008, p. 3) He goes on to outline how each of these three metadata types has a particular function, and is covered by existing standards:

- Descriptive/bibliographic metadata
  This describes the intellectual content of the object, and is used for searching purposes. Typical fields are title, abstract, author, and keywords. Standards include Dublin Core, MODS and MARC.
- Administrative metadata.
  This provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. There are various subsets of administrative data; two that sometimes are listed as separate metadata types are:
  - Technical metadata
    This provides details of the physical characteristics of a digital object (OCLC/RLG, 2005, pp. 2-3) and can include, for example technical metadata for still Images (standard MIX[4]), text (standard TEI[5]), audio (standard AUDIOMD[6]), and video (standard VIDEOMD[7]).
  - Rights management metadata (ODRL[8]) (OCLC/RLG, 2001b) which deals with intellectual property rights pertaining to the digital object.
- Structural metadata
  This indicates how compound objects are put together, for example, how pages are ordered to form chapters. METS is a standard for structural metadata that is widely accepted within the library community.

Other descriptors for metadata have entered the field, such as preservation metadata; an umbrella term used in the PREMIS standard to denote "the information a repository uses to support the digital preservation process" (OCLC/RLG, 2005, p. ix). Preservation metadata thus covers administrative, technical and structural metadata, highlighting the somewhat fluid nature of definitions in this field that make it difficult to consistently draw clear boundaries around different kinds of metadata.

3.2.2 Preservation Metadata Roles / Approaches

There are differing roles played by preservation metadata.

According to Lavoie and Gartner, the categories of information required in a digital preservation context include (Lavoie & Gartner, 2005):

- provenance (custodial history of the object)
- authenticity (information to validate if the information object is what it purports to be)
- preservation activity (actions taken to preserve the object and results thereof)
- technical environment (what hardware, operating system, and application required to render the object)
- rights management (who can and cannot access the object)

Woodyard-Robinson suggests that metadata supports the functions of an archival information system to maintain (Woodyard-Robinson, 2006):

- Viability: ensures the object remains intact and documents changes if any transformation has been made;
- Renderability: ensuring the bit streams are interpreted and rendered;
- Understandability: ensure digital objects can be understood both by machines and humans;
- Authenticity: validate if the object is what it purports to be
- Identity of digital documents

Rothenberg and Biksen take the view is that "the design of metadata related to the technology of preservation must await further experimentation and prototyping of specific preservation approaches" (Rothenberg & Bikson, 1999, p. 8)

The National Library of New Zealand metadata initiative (*Metadata Standards Framework –Preservation Metadata (Revised)*, 2003, p. 3) identified two major functions of preservation metadata:

- to provide sufficient knowledge to take appropriate actions in order to maintain a digital object's bit stream over the long-term
- to ensure the content of an archived object can be rendered and interpreted, in spite of future changes in storage and access technologies.

Caplan provides a number of examples of preservation activities and indicates how metadata can be used to support them (Caplan, 2009):

- Checksum information stored as metadata can be used to tell if a stored file has changed between two points in time.
- Metadata can support media management by recording the type and age of storage media and the dates that files were last refreshed.
- Metadata about original file formats and the hardware and software environments supporting them can be used to inform migration or emulation.
- Metadata can help support authenticity by documenting the digital provenance of the resource and its chain of custody and authorized change history

3.3 Open Archival Information System (OAIS)

The Reference Model for an Open Archival Information System (OAIS) was first issued in January 2002 as a recommendation (Blue Book) from the Consultative Committee for Space Data Systems (CCSDS).[9] It was subsequently approved as an ISO standard (No. 14721: 2003) (See: CCSDS, 2002) The report marked the culmination of an iterative development process and international consultation activity which had taken around seven years to complete.

The OAIS reference model sets out definitions of a functional archival information system (Lavoie, 2004). It defines concepts and terminologies that are recommended to be used in information systems that are responsible for long-term preservation of information objects.

It is very common[10] for repositories to assert themselves to be 'OAIS-compliant' (Lavoie & Gartner, 2005, p. 9), but it is not always clear what this means in practice. Neither the CCSDS nor the ISO have a formal certification process by which

compliance claims may be independently verified. In order formally to comply with the OAIS model, archival information systems need to fulfil two major criteria:

- the system should fulfil the six OAIS mandatory responsibilities(CCSDS, 2002):
  - negotiating and accepting information from producers
  - having enough mandate on the information
  - determine designated community(OAIS users)
  - ensure understandability and usability of the content
  - using appropriate policies and procedures to ensure authenticity and originality
  - ensuring availability of the preserved information available to the Designated Community.
- They should use OAIS concepts and terminologies (CCSDS, 2002).
  - This is designed to assist interoperability by enforcing common locution. Examples include the use of Designated Community, Submission Information Package (SIP), Archival Information Package (AIP), and Dissemination Information Package (DIP).

The 'OPEN' in 'Open Archival Information System' draws attention to the open public discussion which led to the model's creation. An Archival information system is a combination of people and systems (manual or computer) which has a responsibility of long-term preservation of information resources for a designated community (CCSDS, 2002). The OAIS functional model in turn has six components: ingest archival storage, data Management, preservation planning, access and administration (CCSDS, 2002).

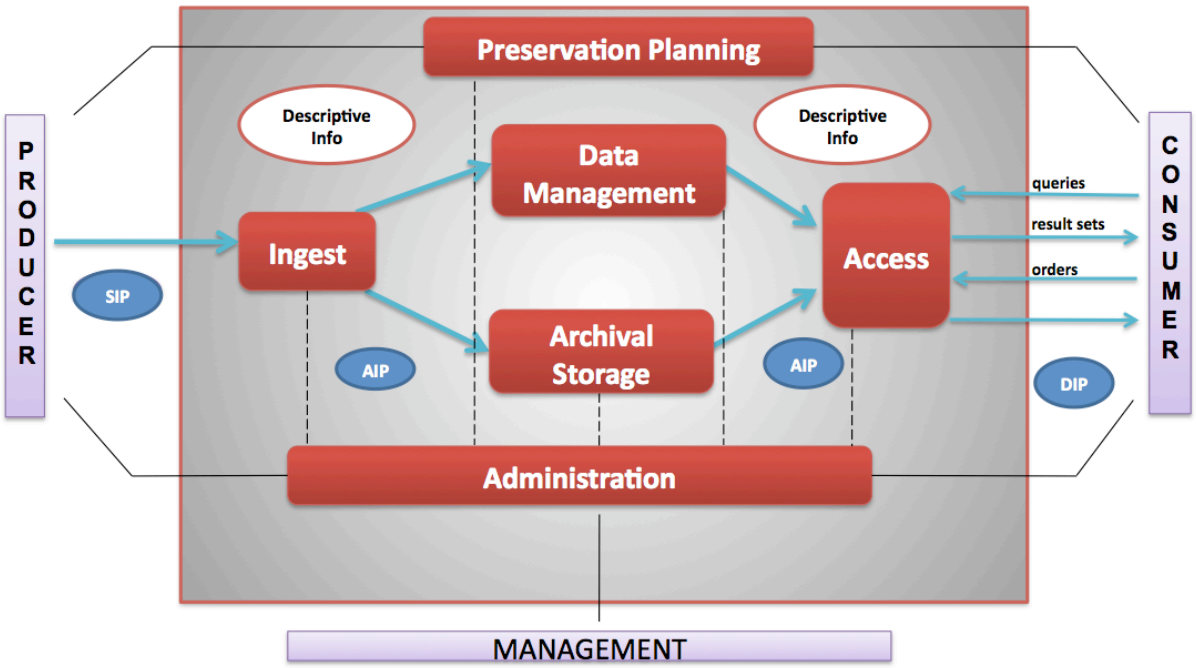

Fig.1. The OAIS Functional Model
(Based on: CCSDS, 2002).

As indicated in the functional model, OAIS has an environment with which it regularly interacts in order to achieve its goals. Producers of the information produce and

submit content with metadata to be ingested in the archive, consumers of information send requests to the archival system and use the result retrieved from the system. OAIS refers to its users as a 'designated community'. Management (which is responsible for the strategic operation of the archive) is also one of the actors within the OAIS environment (Lavoie, 2004).

Within its boundaries, OAIS has two major components: the functional model and information model. The information model is particularly relevant to describe the metadata requirements for long-term preservation (OCLC/RLG, 2002, p. 5).

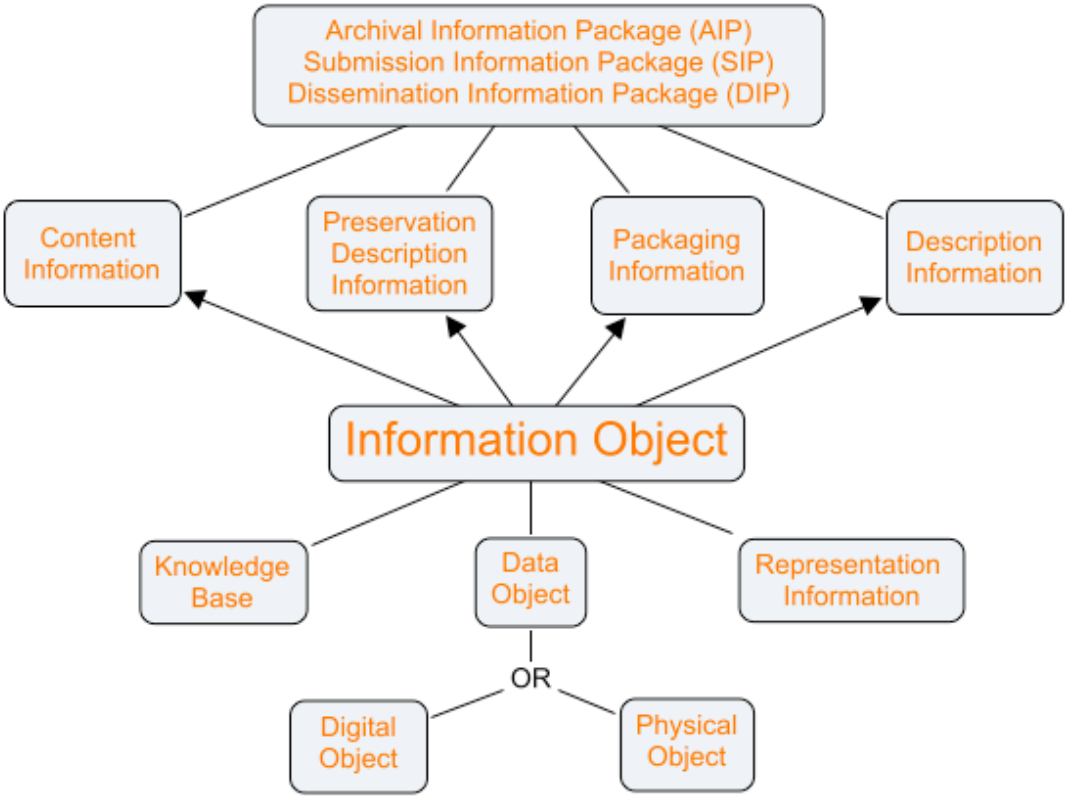

Fig.2. The OAIS Information Model (Based on:OCLC/RLG, 2002, p. 6).

The OAIS information model has the information package and associated metadata. There are three types of information packages: the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP) (CCSDS, 2002; OCLC/RLG, 2002).

According to Rothenberg there are problems with the OAIS model:

"By offering a standardized terminology and viewpoint, the OAIS has facilitated coherent dialogue and cross-fertilization of ideas surrounding digital storage and access across such diverse communities as archives, libraries, and data warehouses. Yet despite the ubiquitous use of the term 'archival' in its documentation, the OAIS model is relatively silent on the subject of preservation per se: it focuses on the processes of describing, packaging and manipulating stored information with relatively little discussion of how to keep it meaningfully readable and usable in the future." (Rothenberg, 2000)

The OAIS model claims to be independent of any particular preservation strategy (CCSDS, 2002), but according to Rothenberg:

"Because it focuses only superficially on the technical issues involved in making digital information readable and understandable in the far future and because it implicitly assumes that migration will be the only method used to preserve such information, the OAIS presents an 'archival' model that ironically says very little about preservation." (Rothenberg, 2000)

In relation to metadata, the model includes Preservation Description Information (PDI), i.e. information about the unique reference number, provenance (who does what and when) and fixity information (whether the object has been changed).

Since OAIS is a conceptual framework, it is considered as a theoretical conception of an archival information system (manual or digital) that comprises of the information model (content plus metadata). As a conceptual model, OAIS informs the development of archival information systems, tools and services. It is not by itself an implementable model.

3.4 Preservation Metadata Projects and Initiatives

Some national libraries, archives and museums are required by law to ensure long-term access to their nation's cultural heritage. Information and communication technologies have created unprecedented opportunities of generation, creation and dissemination of digital information to which institutions have responded by investing in a variety of preservation projects. The following sections provide brief overviews of some of the major digital preservation data models and metadata initiatives and projects.

3.5 The Research Libraries Group (RLG)

In 1998, the Research Libraries Group of the Library of Congress developed a set of 16 metadata elements for digital images preservation:

- Date
- Watermark
- Transcriber
- Resolution
- Producer
- Compression
- Capture
- Device Source
- Capture
- Details Color
- Change
- History
- Color Management
- Validation Key
- Color Bar/Grayscale Bar
- Encryption
- Control Targets

"The RLG elements illustrate the relationship of preservation metadata to the three broad categories of metadata defined above: descriptive, administrative, and structural. Although preservation metadata can potentially straddle all three metadata types, its focus lies with the latter two." (OCLC/RLG, 2001a, p. 5).

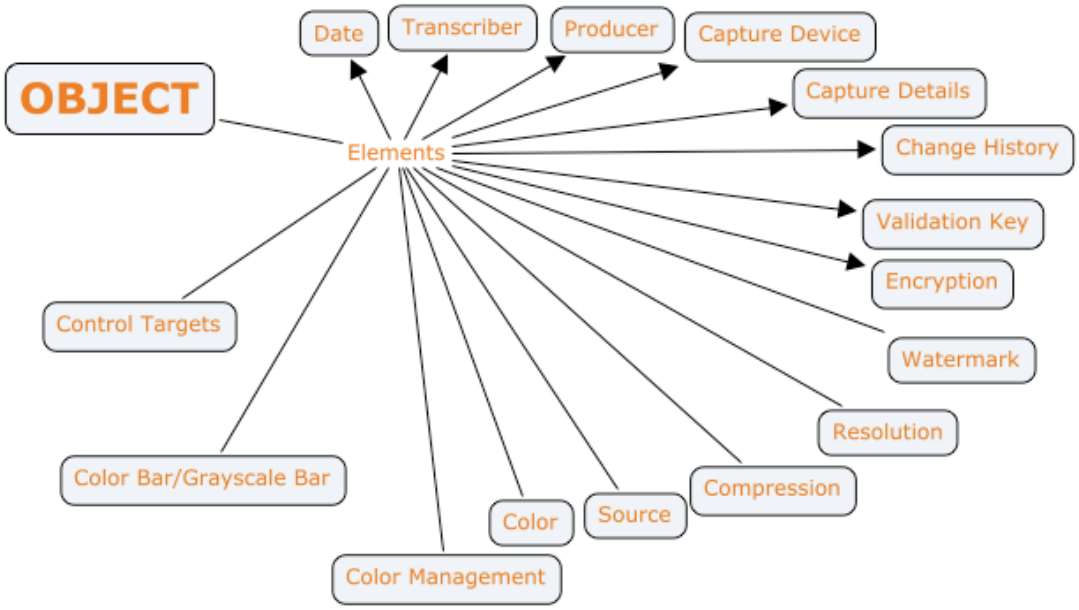

Fig.3. RLG metadata elements

The RLG's metadata set was aimed at facilitating the preservation of and access to digital images which makes it of limited use for other types of digital objects. This was a preliminary attempt by the RLG to define metadata schemas for digital images. It was not implemented widely but it helped reinforce the discussions to work on preservation metadata not just for images but in general for digital objects.

3.6 PANDORA Logical Data Model (National Library of Australia)

The PANDORA project was initiated by the National Library of Australia (NLA) in 1996 to help ensure long-term access to significant Australian on-line publications.

The aims of the project were to:

- identify the functional requirements for a management and preservation system
- identify, test and evaluate a range of technical methods, standards and products involved in the preservation process, including capture, cataloguing and archiving.
- estimate the financial, equipment and staff resources required for the ongoing management of publications selected for national preservation
- develop recommendations and strategies for the long term preservation of and access to publications selected for national preservation, including considerations of data refreshment, reformatting and migration to counter obsolescence or loss
- develop a proposal for a national approach to the long-term preservation of these publications (Source: NLA, 1997)

The following diagram shows the various interrelated components of the PANDORA data model.

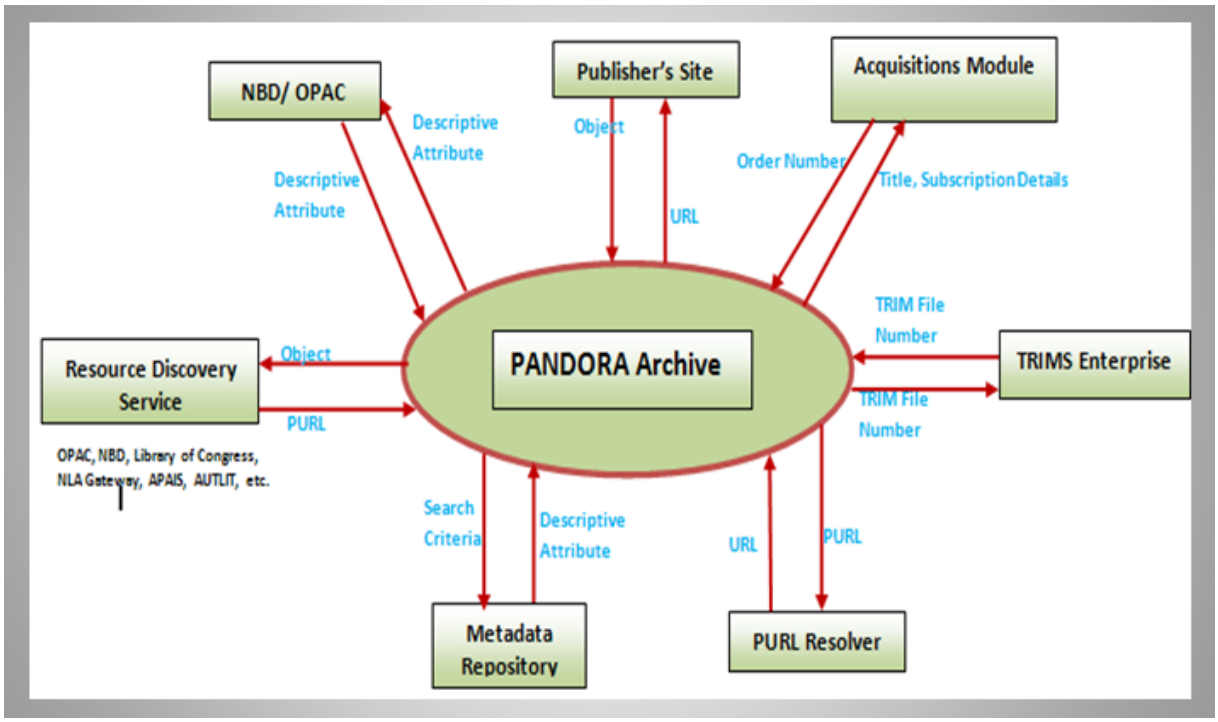

Fig.4. PANDORA Data Model (Based on: NLA, 2001)

In 2001, NLA revised PANDORA and developed PANDAS (PANDORA Digital Archiving System) to serve as an integrated, web-based, web archiving management system. Version 2 of PANDAS was released in 2002 and Version 3 in 2007.

PANDAS was designed to support workflows including:

- identifying, selecting and registering candidate titles
- seeking and recording permission to archive
- setting harvest regimes
- gathering (harvesting) files
- undertaking quality assurance checking
- initiating archiving processes
- organizing access, display and discovery routes to, and metadata for, the archived resources (NLA, 1999)

According to Woodyard & Bradley, PANDORA's draft preservation metadata framework was informed by the OAIS model and acquired experiences from other metadata schemas such as Cedars, NEDLIB, The Making of America II Project, the National Archives of Australia's Recordkeeping Metadata Standard and RLG (Woodyard & Bradley, 2000).

The NLA metadata schema has the following elements:

- Persistent Identifier
- Date of creation
- Structural type

- Technical Infrastructure of Complex Object
- File description
- Known System Requirements
- Installation Requirements
- Storage Information
- Access Inhibitors
- Finding and Searching Aids and Access Facilitators
- Preservation Action Permission
- Validation
- Relationships
- Quirks
- Archiving Decision (work)
- Decision Reason (work)
- Institution Responsible for Archiving Decision (work)
- Archiving Decision (manifestation)
- Decision Reason (manifestation)
- Institution Responsible for Archiving Decision (manifestation)
- Intention Type
- Institution with preservation responsibility
- Process
- Record Creator

(Woodyard & Bradley, 2000) (Day, 2001)

The extent of the uptake of PANDORA and PANDAS across Australia is not clear from the literature reviewed although survey work done by the PREMIS Working Group suggests that schemes coming from the NLA did not have any significant international take-up (OCLC/RLG, 2004)[11].

3.7 Preservation Metadata Standards Framework (National Library of New Zealand)

The first version of the National Library of New Zealand (NLNZ) metadata schema was released in November 2002. The aim was to support the digital preservation activity of the NLNZ. The metadata schema was revised in 2003 . The initiative initially identified preservation metadata elements for:

Object

- Name of object
  - Reference number
  - Identifier - Object IID
  - Persistent Identifier – PID
  - UNIX location
  - Date of creation of Preservation Master
  - Technical composition
  - Structural type
  - Hardware environment
  - Software environment
  - Installation requirements

    - Access inhibitors
    - Access facilitators
    - Quirks
    - Authentication
    - Metadata record creator
    - Date of metadata record creation
    - Comments
- Process
    - Object IID
    - Process
    - Purpose
    - Name of the Individual, Business, Unit or Agency
    - Permission
    - Date of permission
    - Hardware used
    - Software used
    - Steps
    - Result
    - Guidelines
    - Completion date and time
    - Comments
- File
    - Object IID
    - File IID
    - Structural context
    - Filename and extension
    - File size
    - File date and time
    - MIME type/format
    - Version
    - Target indicator
- Image
    - Resolution
    - Dimensions
    - Tonal resolution
    - Colour space
    - Colour management
    - Colour lookup table
    - Orientation
    - Compression
- Audio
    - Resolution
    - Duration
    - Bit rate
    - Compression
    - Encapsulation
    - Track number and type

- Video
  - Frame dimension
  - Duration
  - Frame rate
  - Compression
  - Video Encoding Structure
  - Video Sound
- Text
  - Compression
  - Text Character set
  - Text Associated DTD
  - Text Structural Divisions
- Data Sets
  - No unique fields
- System Files
  - No unique fields

(Source: NLNZ, 2003)

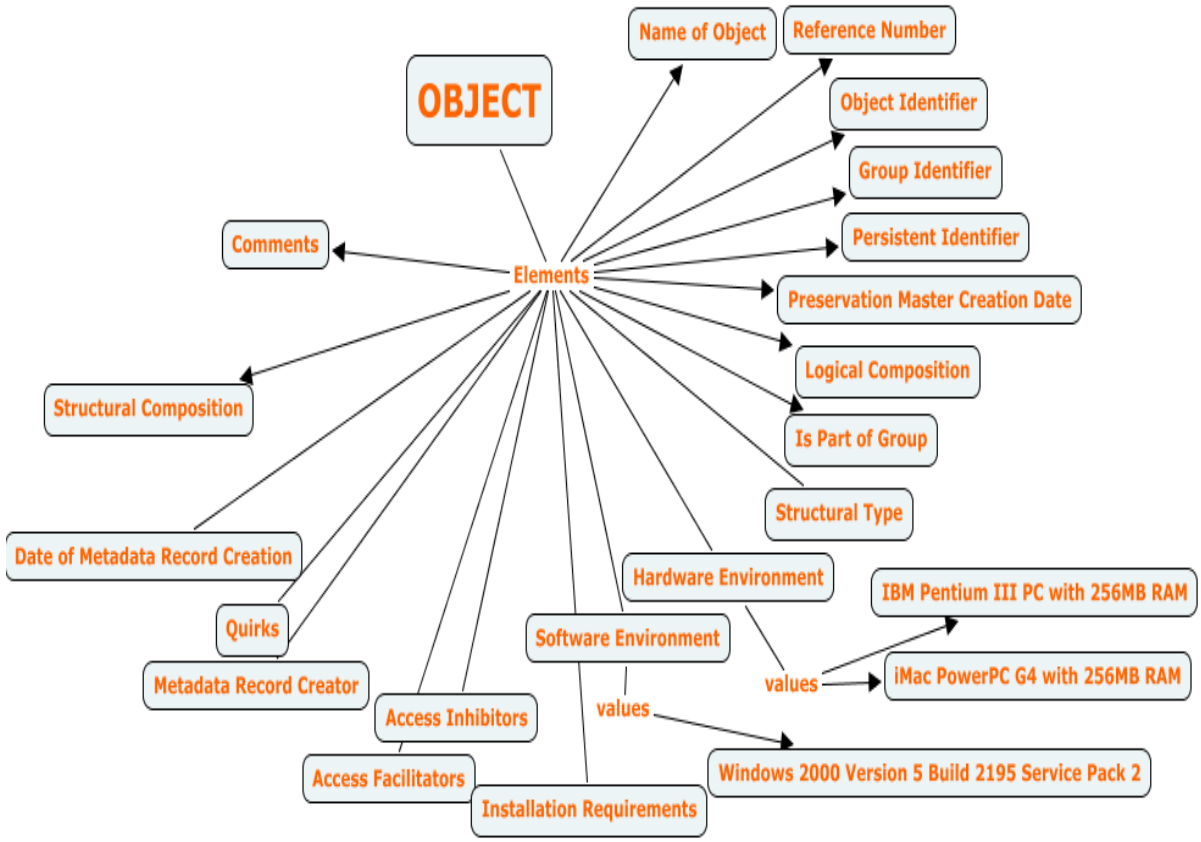

Fig.5. NLNZ metadata schema (partial) for the 'Object' metadata element

The NLNZ schema was seen as significant in virtue of having been one of the first

preservation metadata schemas that was actually implemented. As a result the data model of the NLNZ schema inspired the German development of LMER in 2003. The NLNZ schema includes information about hardware and software environments and also includes information about rights and provenance. The schema recognized the possibility of future changes and revisions to comply with other international standards (NLNZ, 2003).

3.8 Networked European Deposit Library (NEDLIB) Metadata Elements

The NEDLIB project defined a functional model based on the OAIS reference model. The functional model is called Deposit System for Electronic Publications (DSEP). The NEDLIB DSEP model adapted functions of OAIS: ingest, archival storage, data management, access, and administration (van der Werf-Davelaar, 1999). The OAIS function 'preservation planning' was subsequently introduced as an input of NEDLIB (van der Werf, 2000). The DSEP data model includes the original bit stream of digital publications, metadata, software, and packaging information. The DSEP model stores and manages metadata separately from the digital object (bit stream). The reason for this is that while the bit stream does not change, the metadata for it may be changed frequently (Day, 2001).

NEDLIB has metadata elements and sub-elements arranged as follows (Lupovici & Masanès, 2000):

Representation Information

- Specific Hardware requirements
  - Specific microprocessor requirements
  - Specific multimedia requirements
  - Specific peripheral requirements
- Operating system
  - Name
  - Version
- Interpreter and compiler
  - Name
  - Version
  - Instruction
- Object Format
  - Name
  - Version
- Application
  - Name
  - Version

Preservation and Description Information

- Reference Information
  - Creator
  - Title
  - Date of Creation
  - Publisher

    - Assigned Identifier
        - Value
        - Construction method
        - Responsible agency
    - URL
        - Value
        - Date of Validation
- Fixity Information
    - Checksum
        - Value
        - Algorithm
    - Digital Signature
- Change History
    - Main metadata concerned
        - Date
        - Old Value
        - New Value
        - Tool
        - Name
        - Version
    - Reverse
    - Other metadata concerned
        - Old Value
        - New Value

The metadata element categories correspond with OAIS terminologies such as Representation Information (Specific Hardware requirements, Operating system, Object format, Application, etc.); PDI and Descriptive Information (Reference Information, Assigned Identifier, URL, Checksum, Change History, etc.). Lupovici & Masanès indicate for each metadata element or sub-element its relationship to the NLA and CEDARS schemas, however the extent of the overlap is not very substantial (Lupovici & Masanès, 2000, pp. 21-25).

In a DSEP environment various 'technical' characteristics of a publication are held. "In particular:

- The **format(s)** of a publication, referring to the way in which the data is encoded (file formats, character encoding, etc.).
- The **navigational structure** of a publication, referring to directory/files structure of a publication (table of contents, numbered list of items, navigational structure with hyperlinks, etc.).
- The **application software** accompanying the publication, referring to the software that is required to "render" the publication (viewer, browser, search and retrieval software, etc.).
- The **system requirements**, referring to the hardware and systems software configurations that can run the publication." (van der Werf-Davelaar, 1999).

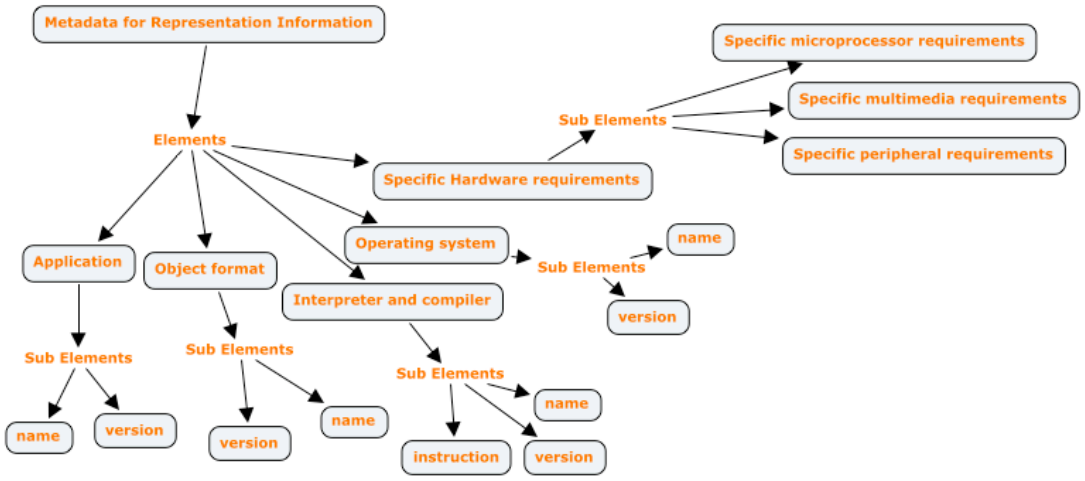

Fig.6. NEDLIB Metadata Elements (Representation Information)

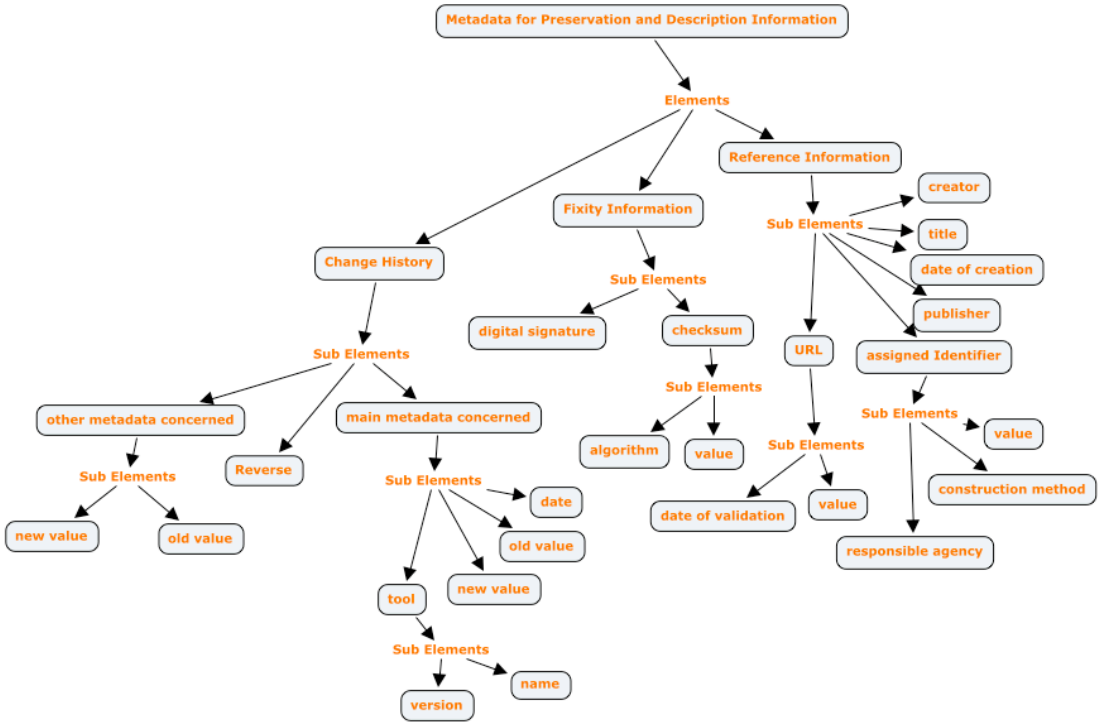

Fig.7. NEDLIB Metadata Elements (Preservation and Description Information)

The NEDLIB project identified the centrality of the authenticity of the digital original and the project recognized the challenges of identifying what needs to be preserved and emulated, and paid particular attention to the need to specify metadata in sufficient detail to enable future generations to be able to access and render the digital publications (van der Werf-Davelaar, 1999).

While NEDLIB's metadata schema was a significant early effort in preservation metadata, there is little evidence in the literature that NEDLIB was widely adopted by other libraries and archives; for example the PREMIS Working Group reported that just 15% of the institutions it surveyed in late 2003 were using or 'planning to use' NEDLIB (OCLC/RLG, 2004). Nevertheless, NEDLIB did contribute to the development of other metadata schemas including PREMIS. Since NEDLIB includes hardware and software dependency metadata elements, it represents an important foundation on which further development in this area might be built.

3.9 CEDARS Preservation Metadata Elements

CEDARS (CURL Exemplars in Digital ARchiveS) arose out of Phase 3 of the JISC-funded Electronic Libraries (eLib) Programme which promoted research into some of the practical aspects of digital preservation (Day, 1998). It was a collaborative effort involving UKOLN (The UK Office for Library and Information Networking) and CURL (Consortium of University Research Libraries[12]).

The aims of CEDARS were to:

- Promote awareness about the importance of digital preservation amongst
- research libraries and their users
- the data-creating and data-supplying communities
- Identify, document and disseminate strategic frameworks within which individual libraries can develop collection management
- Safeguard the long-term viability of any digital resources that are held in libraries by helping to guide decision-making
- Investigate, document and promote methods appropriate to the long-term preservation of different classes of digital resources typically included in library collections
- Develop costed and scaleable models

The lead sites in the project were the Universities of Cambridge, Leeds and Oxford. UKOLN was involved with the metadata aspects of CEDARS . Other collaborating institutions included the Arts and Humanities Data Service (AHDS), the British Library, the Data Archive, the NPO and the Research Libraries Group (RLG).

One of the results of the CEDARS project was the development of the CEDARS Preservation Metadata Elements.

The following metadata elements were identified by CEDARS:

Content Information

- File Identifier
- File Size

Representation Information

- Language
- Storage format
    - Type
    - Version
    - Specification
- Name of File Type
- Version of File Type
- Number of files
- Total file size
- Documentation
- File Modality
- Character Set
- Associated Data Type Definition

- Encoding
- Resolution
- Colour
- Colour Management
- Colour Bar / Grey Scale Bar
- Sampling Frequency
- File Encoding
- Bits
- Mono/Stereo
- Application-Dependency
- Application Name
- Application Version
- Notes
- Software Environment Dependency
- Software Environment Name
- Software Environment Version
- Hardware Dependency
- Hardware Name
- Hardware Type
- Hardware Description
- Specification
- Modality Served
- Compression
- Name (of compression used)
- Method
- Required uncompression package
- Location

Preservation Description Information: Context Information

- Identifier Name
- Identifier Value
- Level of granularity
- Related item
- Relation to item
- Is item required?
- History
    - Submission Date

Change History

- Formatting Change
    - Date
    - Previous format
    - Process
    - Reason
    - Conversion Agent

Preservation Strategy History

- Preservation Strategy

- Removal Authority
- Retention Policy Citation
- Retention Authority Issuance
- Retention Period End Time
- Selection Status

Usage
- Use-Type
- Use Instance Time
- Use Instance User

Format Processing
- Process Type
- Process Agent
- Publisher Preferred Browser

Preservation Description Information: Provenance Information

Publisher Information
- Publisher Name
- Publisher Contact Details
- Rights Management
- Access Profile
- Access Rights

Capture Procedure
- Capture Procedure Type
- Gathering Frequency
- Gathering Regime

Rights information
- Event Type
- Negotiation Status
- Copyright Warning
- Publisher Copyright Statement
- User Class
- Restriction Status
- Access Conditions
- Use Conditions

Preservation Description Information: Fixity Information
- Authentication type
- Authentication key
- Authentication value
- Authentication documentation

Packaging Information

N.B. These choices are the subject of local archive definitions or conventions.

Descriptive Information

- Title
- Author/Creator
- Description
- Contributor
- Date
- Language

(Source: Stone & Day, 1999)

The list while being fairly extensive is, in a few respects, quite loose and occasionally lacks clear definitions of the elements and what kind of values they accept. A case in point is the element 'File Size' which, in the accompanying comments, lacks information on whether it is repeatable and where the value should come from. Some of the element names are inadequately descriptive, e.g., 'Bits' (representation information for the number of bits an audio sample contains).

The CEDARS metadata specification defines basic information about Environments such as Hardware Dependency, Environment (OS) Dependency and Application Dependency. See figure below.

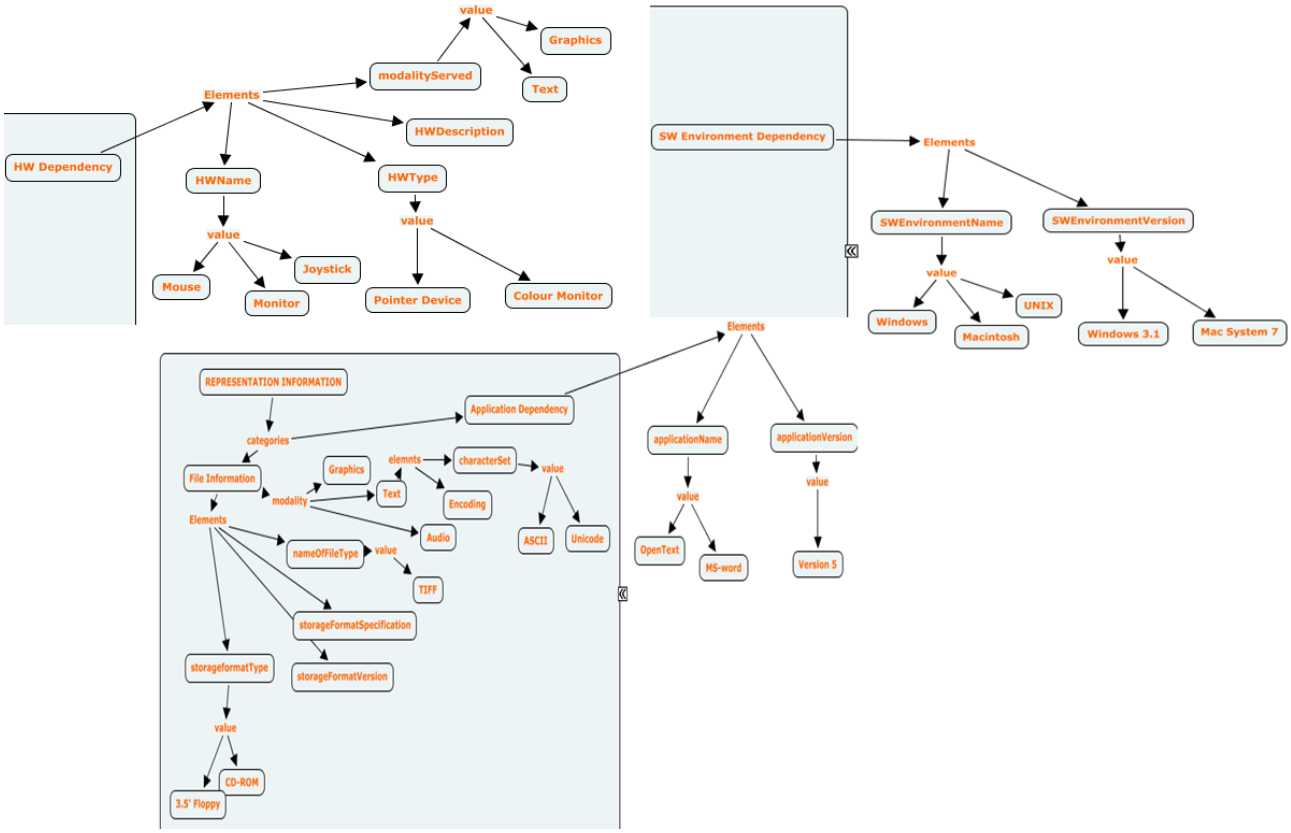

Fig. 8. CEDARS Metadata Elements (Based on: Stone & Day, 1999, p. 2)

CEDARS is a creditable early effort in developing a metadata schema for digital preservation, but its implementation is not visible. The extent to which CEDARS was adopted by libraries was not clear in the literature consulted for this report but perhaps some indication may be given by the PREMIS Working Group Survey of institutions, which reported that just 15% of those responding were using or 'planning to use' CEDARS (OCLC/RLG, 2004). The project, however, implemented a pilot

digital archive to test the metadata architecture. The CEDARS demonstrator was developed based on OAIS.

A strength of CEDARS is its foundation on the OAIS reference model (See Fig. 7 below). The CEDARS team analysed earlier metadata schemas such as PANDORA, the Pittsburgh project, the RLG Working group on Preservation Issues of Metadata (Source: Stone & Day, 1999)

It uses Dublin Core for its descriptive information.

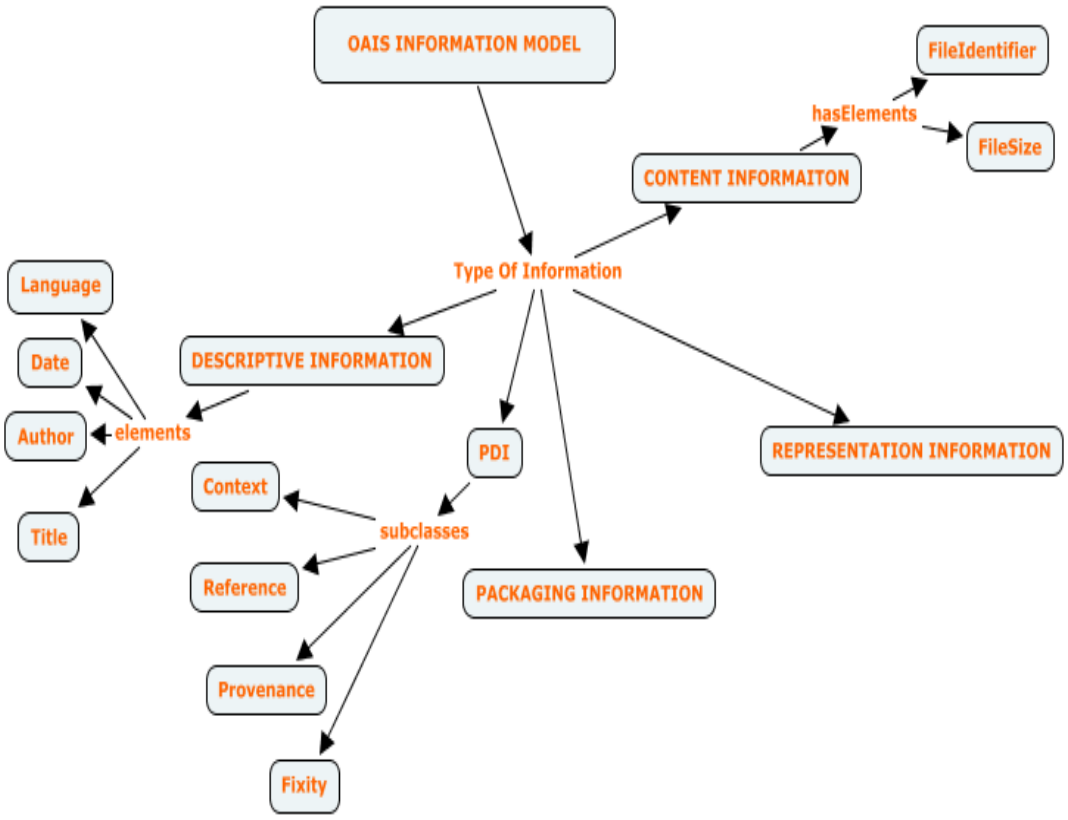

Fig. 9. CEDARS metadata elements

According to the CEDARS summative evaluation report (Owen, Coutts, & Veling, 2001)

“One of the most important deliverables of the CEDARS project would seem to be the draft outline specification of metadata for digital preservation. This document is has attracted international interest as it offers a translation of the abstract OAIS model into practical metadata specifications”.

CEDARS certainly had some influence on the work of other groups. Maggie Jones reports that “The work of RLG and OCLC in attempting to arrive at a standard set of preservation metadata elements, made use of the Cedars outline specification in its deliberations.” (Jones, 2002, p. 14)

While from the perspective of the OCLC/RLG the influential role of CEDARS is equally clear: “The CEDARS preservation metadata element set was drafted for two purposes: first, as an implementable scheme for use with the CEDARS pilot digital

archive, and second, as a contribution toward an international effort to cooperatively develop a standard set of preservation metadata elements." (OCLC/RLG, 2001a, p. 17)

According to the CEDARS evaluation report : "the project has been above all a theoretical exercise, with an implementation of the theoretical concepts in a proof-of-concept demonstrator archive, but remains to be validated in the context practical applicability of the guidelines and of other work on preservation strategies and mechanisms." (Owen et al., 2001)

It is perhaps worth mentioning in conclusion that CEDARS was not oriented towards practice but was focused on theoretical and methodological issues of digital preservation. The approach towards metadata within the KEEP is pragmatic and every effort will be made to pay attention to the practical needs of librarians and archivists as well as end-users.

3.10 PREMIS (PReservation Metadata Implementation Strategies)

PREMIS is the result of a concerted effort of an international working group which was set up by the Online Computer Library Centre (OCLC) and the Resource Libraries Group (RLG) in 2003. The group included thirty international experts from the US, the Netherlands, Australia, Great Britain, Germany, New Zealand and from the CEDARS project team (Lavoie, 2004). The working group's 2005 report on preservation metadata was considered as a significant milestone in the search for a common implementable metadata framework for preservation. The PREMIS working group "was charged to define a set of semantic units that are implementation independent, practically oriented, and likely to be needed by most preservation repositories" (Caplan & Guenther, 2005, pp. 111-112).

PREMIS defines preservation metadata as "the information a repository uses to support the digital preservation process" (OCLC/RLG, 2005, p. ix) and considers metadata essential to maintain viability, renderability, understandability, authenticity, and identity. From the PREMIS perspective, preservation metadata is not limited to technical metadata but includes different categories of metadata including rights metadata and provenance metadata.

Technical metadata is described as "the physical rather than intellectual characteristics of digital objects."(OCLC/RLG, 2005, p. x) The group appreciated that detailed, format-specific technical metadata is important to implement preservation strategies but nevertheless did not specify any technical format-specific metadata for digital objects, preferring to leave "the definition of metadata for describing media and hardware characteristics to specialists in these areas".(OCLC/RLG, 2005, p. 3). Without this kind of information an emulation platform cannot be successfully implemented.

Environment metadata is defined by PREMIS as "application software, operating systems, computing resources and even network connectivity [that] allows the user to render and interact with the content".

'Semantic unit' is the PREMIS term for what is referred to in other metadata standards a 'metadata element' Semantic units are the properties of entities, so, in database terminology, they are attributes. A semantic unit can group together other semantic components (i.e. it can be an umbrella term. (OCLC/RLG, 2005, pp. 1-1; 1-2; 1-4).

- An object is defined as "discrete unit of information in digital form" and can be a representation, a file and/or a bit-stream".

- A file is a “named and ordered sequence of bytes that is known by an operating system”.
- A representation is a single digital instance of an intellectual entity held in a preservation repository.
- A bit stream is the sequence of bits that make up a file. (Woodyard-Robinson, 2007, p. 51).

This distinction is important for digital preservation as it addresses the fact that a digital object often includes other digital objects embedded within it. For instance a web page might include several objects within itself. Alternatively a thesis (intellectual entity) could contain three PDF files. Each PDF file might contain several images (bit streams). .

The approach used by PREMIS; treating a digital object at three different levels, is therefore important.

Since emulation relies heavily on technical and environment metadata, it is imperative that libraries, museums, and cultural heritage institutions keep metadata about the technologies in which their preserved digital objects were created and on which they were accessed. The granularity required of such metadata depends greatly on the emulation platform. If the emulation is done at a hardware level, well-structured information (metadata) about both the operating system and the application software necessary to render the objects must be recorded.

The PREMIS working group recommends the use of controlled vocabularies for preservation metadata values. It further recommends having a central registry of environments metadata which can be shared by different users (OCLC/RLG, 2005, pp. 4-2). Some libraries use PREMIS in combination with format registries and characterisation tools such as PRONOM/DROID, and JHOVE.

It should be noted at this point that however many opportunities a metadata schema or registry may afford preservation professionals to capture information vital for future emulation or migration, it is not always the case that capture actually takes place. For example in Fig.10 (below) it can be seen that no ‘Technical Environment’ details are held by PRONOM on the very common Microsoft for Macintosh 6.0 file format. This is not untypical of PRONOM entries and illustrates something of the gap that needs to be addressed between theory and practice in digital preservation.

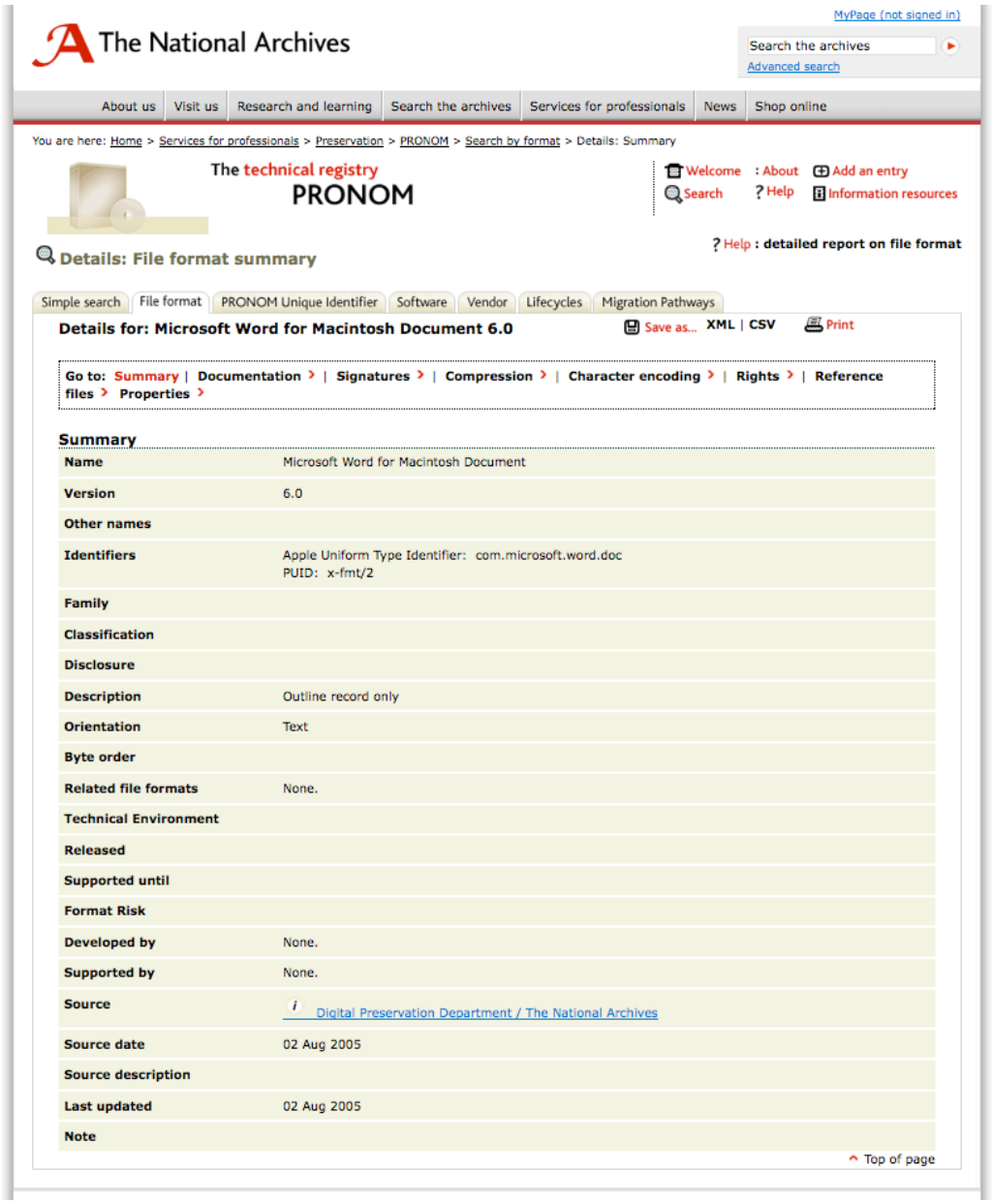

Fig. 10. Details held by PRONOM on the Apple Macintosh Microsoft Word 6.0 file format.[13]

3.11 PREMIS implementation

Compared to other preservation metadata schemas, PREMIS is in the happy position of being widely implemented by libraries and archives. The PREMIS Implementation Registry lists institutions implementing PREMIS:

- The National Archives of Sweden (PREMIS is used for processing and storing digital objects in a digital repository. The National Archives is developing a transfer model for digital objects created in our scanning projects. A function is being developed for packaging and storing data about the digital objects in our archival information system ARKIS partly stored as PREMIS-metadata. The application is in use for storing data. An application for exporting PREMIS data as XML will be developed in the future. Operated in conjunction with a tool called ESSearch) ;

- National Archives of Scotland (The NAS is preparing for the ingest of digital objects from the Scottish Executive[14] and the Scottish Courts. An application is under development that aims to be compliant with OAIS, PD0008 and PREMIS to meet this requirement.);
- Ex Libris (Using JHOVE - to extract technical metadata; URN-NBN - to support the German persistent identifier system; METS - to describe complex objects and to be used as a delivery method);
- Florida Center for Library Automation (The FCLA Digital Archive uses a locally-developed software application called DAITSS, which implements most of the PREMIS data elements.);
- Grainger Engineering Library Information Center, University (This project is designing an architecture and building a toolkit for repository interoperability and preservation. The interoperability is based on a set of METS-profiles developed specifically for this architecture. The METS profiles utilize PREMIS objects for technical metadata and PREMIS events for provenance metadata. The initial release of their tool will support the ingest of our METS packages into the DSpace repository; they anticipate that additional repositories will be supported by subsequent versions of the toolkit with support for Eprints currently under development. The Web Archivists Workbench from OCLC also supports their METS profile as an output format. They are also using METS, MODS, JHOVE, Java XmlBeans, various Apache APIs);
- Cornell University and SUB Göttingen (Using a converter to convert LMER preservation metadata to PREMIS; xmlbeans to build Java objects from JHOVE, METS, and PREMIS XML Schemas; JHOVE to capture technical elements that double as preservation elements; YAR to create OAI-PMH provider for the metadata);
- Library of Congress (PREMIS metadata is generated by the NDNP Validation Library. The NDNP Validation Library both wraps and extends JHOVE.);
- Oxford University Library Services John Rylands Library, University of Manchester (Focussing on the papers of contemporary politicians, the Paradigm project is exploring the issues surrounding the preservation of born-digital personal papers to ensure that they continue to be accessible to the researchers of the future. The project is testing various tools, including JHOVE, Fedora and DSpace, and will develop exemplar strategies and best-practice guidelines that will be of use to any institution which collects, preserves and maintains access to private papers.);

(Source: PREMIS, 2007)

In 2007, the Library of Congress commissioned Deborah Woodyard-Robinson to study how institutions implemented PREMIS. She found that institutions use PREMIS in different ways. None implemented PREMIS 'as is' but instead they used different mechanisms. Some use tools such as DROID/ PRONOM, JHOVE, NLNZ metadata extraction tools. It was reported that there are only very few automatic metadata extraction tools from the objects themselves (Woodyard-Robinson, 2007). Regarding representation of the PREMIS metadata, most of the institutions implemented PREMIS using either a relational database management system or XML.

Woodyard-Robinson concluded that there are still too few implementations of PREMIS having reached sufficient maturity to support firm conclusions on exemplary implementation practises. Her report also found that some repositories want to keep environment metadata on an external repository (Woodyard-Robinson, 2007).

3.12 The need to extend PREMIS

In 2004, the PREMIS working group reported the findings of a survey they had conducted for the purpose of gathering information on key aspects of planned and existing preservation repositories for digital materials

The main findings of the survey were as follows:

- Differences between libraries and archives in terms of materials accepted are significant and reflect the difference in mission. All archives accepted electronic records and the majority accepted datasets and audio/video. Libraries showed less support for datasets and audio/video but more support for locally digitized materials and web resources.
- The cultural heritage/memory institutions lack experience in the area of digital preservation in general and preservation metadata in particular.
- Most of these institutions reported that their preservation practices are guided by the OAIS model, even though the various institutions conform to the model at varying degrees.
- The majority of institutions chose more than one strategy for preservation. Most (85%) are offering bit-level preservation. Beyond that, restrictions on submissions, normalization, migration and migration-on-demand are the four most popular strategies, in that order. According to the respondent's future plans, the four most popular strategies, in order, will be migration, normalization, restrictions on submission, and migration-on-demand.
- Emulation is being used now by only 10% of respondents, but that doubles when future plans are considered.
- Most respondents are recording a wide range of types of metadata; more than half are recording elements of rights, provenance, technical, administrative, descriptive, and structural metadata
- For non-descriptive metadata, METS (Metadata Encoding and Transmission Standard) was by far the most commonly used metadata scheme: 64% of libraries, 42% of archives, and 35% of other institutions used, or planned to use, METS. Z39.87 (Technical metadata for digital still images) was widely used by libraries but not others..
- Most institutions record descriptive metadata manually and they automatically extract technical metadata.
- An emerging best practice appears to be to store content data objects in a filesystem or content management system, and store metadata redundantly in a database and with the data objects. Metadata in the database are used by the repository system for operations, while metadata stored with the objects make the objects self- identifying for preservation purposes. In other models, metadata are stored in either a relational database, an XML database, or in a relational database and partially replicated with the objects.

The following appear to be trends in practice that may ultimately emerge as best practices:

- Store metadata redundantly in an XML or relational database and with the content data objects. Metadata stored in a database allows fast access for use and flexible reporting, while storing them with the object makes the object self- defining outside the context of the preservation repository.
- Use the METS format for structural metadata and as a container for descriptive and administrative metadata; use Z39.87/MIX for technical metadata for still images.
- Use the OAIS model as a framework and starting point for designing the preservation repository, but retain the flexibility to add functions and services that go beyond the model.
- Maintain multiple versions (originals and at least some normalized or migrated versions) in the repository, and store complete metadata for all versions. Retention of the original reduces risk in case better preservation treatments become available in the future.
- Chose multiple strategies for digital preservation. There are good reasons to have more than one approach in a developing field.

(Source: OCLC/RLG, 2004).

PREMIS is a very generic and high-level data dictionary which does not address specific preservation strategies. For instance, metadata elements that are pertinent to emulation are not defined. The environment metadata elements are too generic (such as Hardware Name, Software version, File format) and are not explicitly defined. No guidelines are provided to use the environment metadata elements hence it lacks specificity and detail. According to McCargar, “It is left to individual institutions to apply PREMIS and expand on it according to their needs” (McCargar, 2005, p. 2). This implies that PREMIS needs to be adapted to local situations and preservation strategies as it claims to be core metadata schema.

There are three types of relationships between the various entities. Structural, Derivation, and Dependency relationships. Emulation metadata will mainly use the dependency relationship as it will create relationships between objects, software and hardware environments. Structural relationships will also be used. However, since the goal of emulation is to keep the original objects intact, there is no need to use the derivative relationships. This approach mainly works for migration strategies as it involves transformation of the object.

The PREMIS data dictionary has four semantic entities: Objects, Agents, Rights, and Events (see Fig.11 below). The environment entity is added to emphasize that Environment has not been dealt in detail in PREMIS.

Objects are the subjects of preservation and are well defined. The Environments in which Objects are created, stored, transmitted, and rendered are not defined in great detail as required for example in emulation strategies.

PREMIS has environment dependency as a metadata element but it is not defined explicitly. A promising approach for emulation would be to look at the environment separately, to define descriptive, representative information, and packaging information about the environment itself, thus creating a dependency relationship between Objects and Environment.

In other words, the a specific environment for instance an Operating System could be described using its own elements such as OS-identifier, OS-Name, OS-version, etc.

This kind of information can be populated in to the repository using other services such as PRONOM.

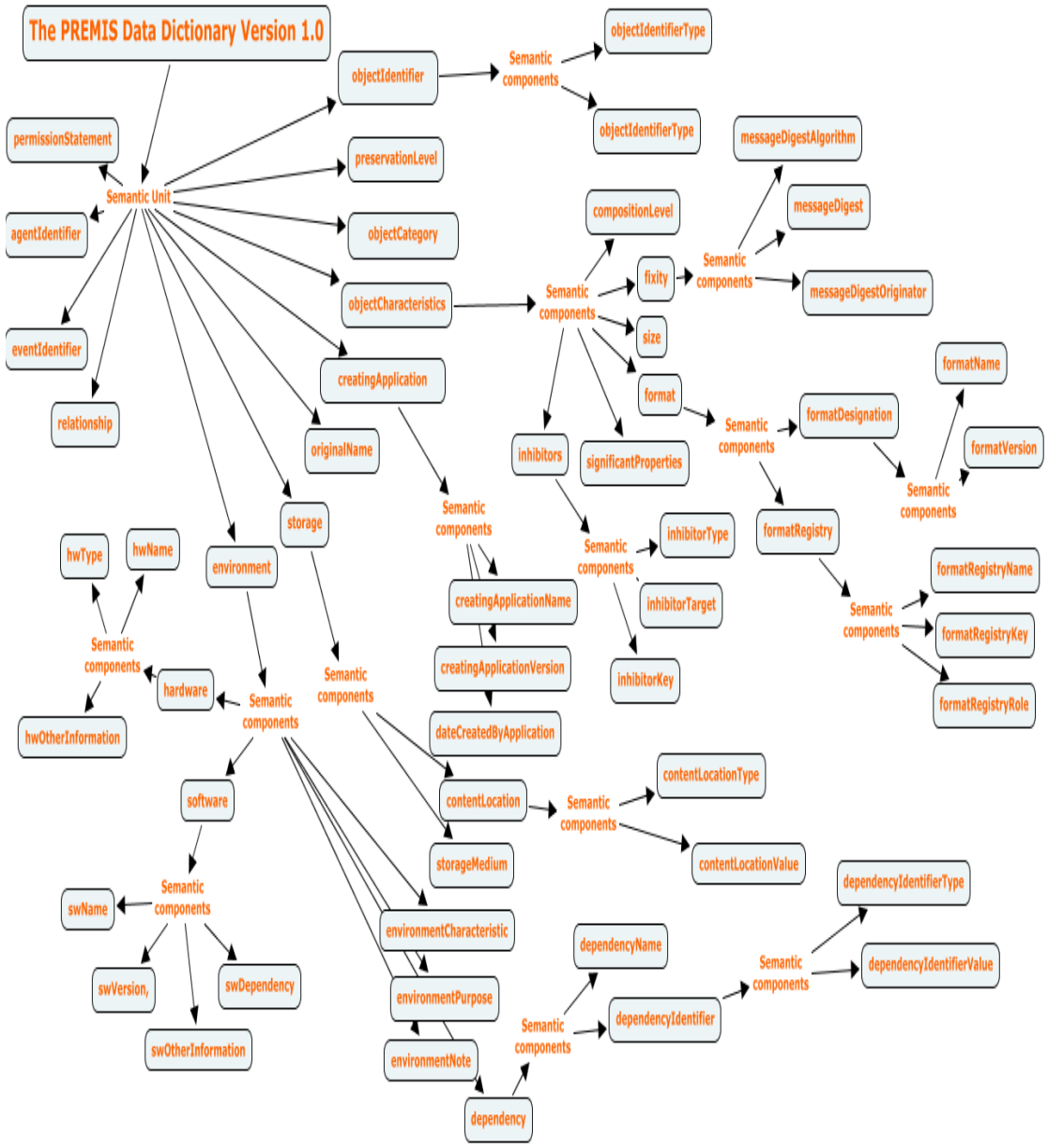

Fig. 11. PREMIS Data Dictionary Source (Based on: OCLC/RLG, 2002)

In PREMIS an effort has been made to use descriptive names which are assumed to help interoperability within different repositories. PREMIS uses containers, which is an approach also used in MARC. But relational databases do not use composite attributes at a single level. Instead the composite attribute is broken down into specific units and each element then has same level of hierarchy with all the other elements. An example of this would be address fields in a database. Dublin core elements are listed without hierarchies.

The metadata elements in the PREMIS metadata dictionary were designed to be general enough to permit individual repositories to define specific metadata elements within their own context and “The working group made no assumptions about the type or structure of the digital resource with which the preservation metadata is

associated, nor did it assume that a particular preservation strategy (e.g., migration or emulation) was followed" (OCLC/RLG, 2002, p. 3).

However, claims of complete generality are somewhat undermined by the relative ease with which PREMIS supports migration strategies compared to the challenges it presents when deploying an emulation approach. This may well reveal that the designers of PREMIS were relying implicitly on an assumption that migration rather than emulation would be the preservation approach employed in practice.

It is important that the environment elements of PREMIS should be extended to permit PREMIS to be used with equal ease for emulation strategies as is currently the case for a migration approach.

3.13 LMER (Long-term preservation Metadata for Electronic Resources)

LMER is a preservation metadata schema developed by the DNB (National Library of Germany). It was designed to address a perceived deficiency in the technical metadata crucial for the long-term preservation of electronic document. LMER was influenced by previous metadata models and schemas and relied heavily on the data model developed by the National Library of New Zealand for their Preservation Metadata Standards Framework[15] LMER is not a general data model for long-term preservation metadata but is designed for concrete use as an exchange format (Steinke, 2005). The LMER schema defines technical metadata, for image, video, audio and text using a modular approach and with an implementation of an XML namespace.

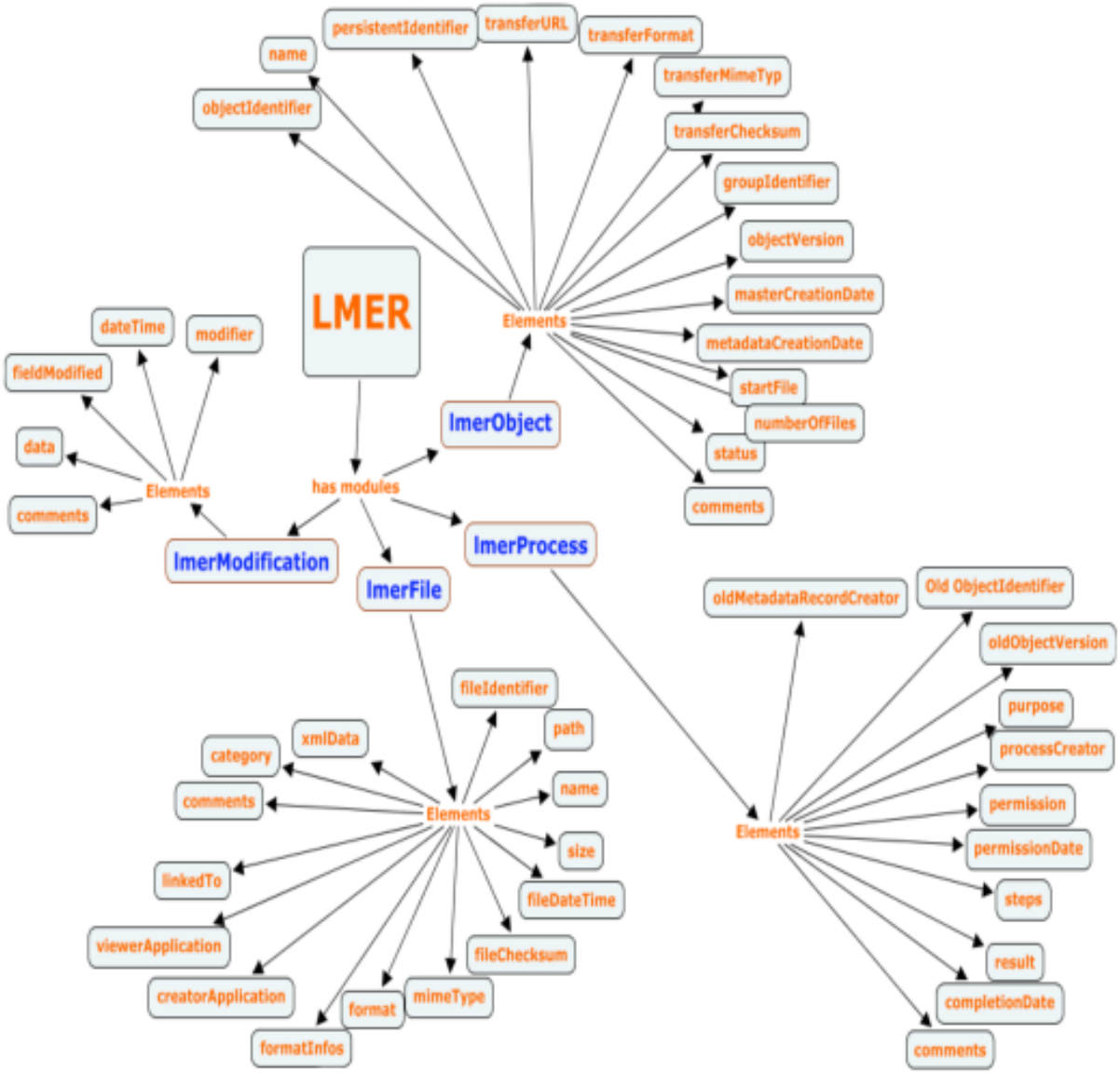

Fig. 12: LMER metadata elements

(Based on: Steinke, 2005)

LMER provides information about hardware and software that is required to render a digital object and is potentially very useful in informing the development of metadata for emulation environments.

3.14 METS (Metadata Encoding and Transmission Standard)

Once the preservation metadata elements have been decided by a digital repository or an archive, the next step is to encode, manage and transmit metadata.

There are two ways in which this process may be managed:

- affixing metadata with the digital object itself
- storing the metadata outside the object

The most frequent means by which metadata storage is accomplished outside the digital object are databases and XML files. One of the XML-based metadata standards that libraries and archives are beginning to use is METS. This was developed in 2001 by the Digital Library Federation (DLF) with support from the Library of Congress. METS can be used to encode descriptive, administrative, and structural metadata using a single XML-based file[16]. METS also facilitates interoperability and metadata cross-walking between different repositories[17].

METS has seven sections (Library of Congress, 2009):

- METS Header
- Descriptive Metadata
- Administrative Metadata
- File Section
- Structural Map
- Structural Links
- Behaviour

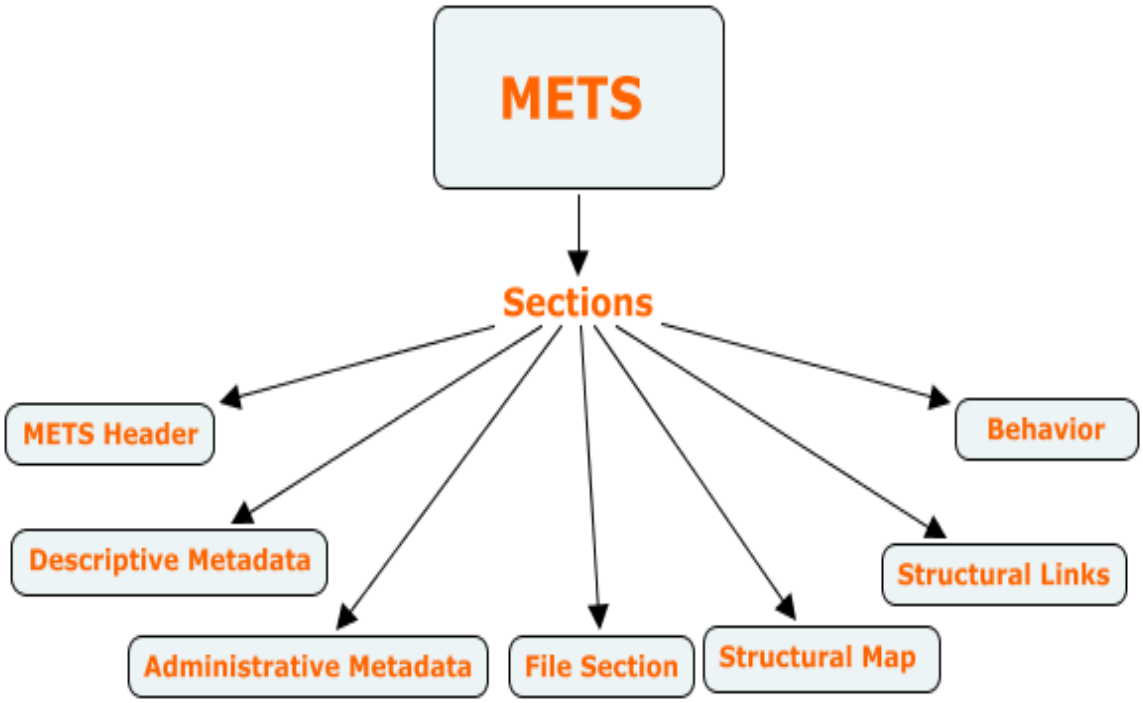

Fig. 13. METS Sections

(Based on: Library of Congress, 2009)

Different types of metadata standards can be used in a METS/XML document. In principle, METS accepts any kind of XML-based data and different types of metadata schemas but, in practice, some are more compatible than others. PREMIS (Version 2) is claimed to be compatible with METS. This, however, requires the schemas to have different sections in order to fit into the METS structure. In other words, to embed different standards in a METS document, the elements of the incoming standard need to be categorised in sections, so that each section can be encoded using METS..

Using METS allows a modularized approach of dealing with different types of metadata schemas. Such categories would be linked using the METS link structure. The various metadata schemas might be either embedded within a METS file or stored externally and linked to METS.

However, this approach may be overly complex as the different schemas coming from various initiatives/standards need to be in conformant with the METS structure. In order to address such problems, the development and implementation of METS profiles is recommended. "METS Profiles are intended to describe a class of METS documents in sufficient detail to provide both document authors and programmers the guidance they require to create and process METS documents conforming with a particular profile" (Source: Library of Congress, 2007). A METS profile is a template developed and made available by institutions through best practises and approved by the METS Editorial Board. The profile explains how the METS sections should be used. The profile may also provide an example entry. The purpose of the profile is to ensure interoperability of different archival systems.[18]

Some schemas are already compatible with METS. For example, MODS and Dublin Core can easily go to the descriptive metadata section of the METS XML structure (Gartner, 2008). Because many different schemas and standards may be combined into one METS structure, it is entirely possible to have duplication of metadata elements and values. This requires a decision to be made about whether to store redundant values or to avoid duplication of data during ingest and metadata management (Gartner, 2008).

Since digital preservation metadata spans different types of metadata, the use of METS or METS-like structures to encode such diverse sets of metadata fields is necessary. Such use is essential for the metadata requirements of emulation strategies. This is because, emulation frameworks will require a rich set of detailed administrative, technical, environment and descriptive metadata elements in order to render the object over the long-term without risking losing authenticity.

The British Library's eJournal System uses METS to encode PREMIS (for preservation metadata) and MODS (descriptive metadata) (Dappert & Enders, 2008). The METS implementation registry lists 40 METS projects which implement METS. The list includes the German National Library's Co-operative Development of a Long-Term Digital Information Archive (kopal), the National Library of Wales, Virtual Library of Historical Press (Spain), Oxford Digital Library, Dspace, Harvard University Library, Stanford University Library, University of California, Berkeley and many other national and university libraries, and digital archives.[19]

Due to the open and non-proprietary nature of XML, METS has been adopted up by many institutions as an encoding and transmission standard.

4 Relevant Framework 7 Projects

There are a number of projects, other than KEEP, which are currently looking at problems and challenges in the area of digital preservation. Over the next few sections of this report a very brief overview is provided of the work being supported by the European Commission under Framework 6 and Framework 7 as well as a number of other activities. All of these projects touch directly or indirectly on environment metadata for complex digital objects.

4.1 LiWA[20] - Living Web Archives.

LiWA will develop and demonstrate web archiving tools able to capture content from a wide variety of sources, to improve archive fidelity and authenticity and to ensure long term interpretability of web content. In particular, web archiving faces many of the same challenges as the emulation community, and synergies could be developed across these areas.

4.2 PARSE.Insight[21] (Permanent Access to the Records of Science in Europe)

PARSE.Insight is a two-year project. It is concerned with the preservation of digital information in science, from primary data through analysis to the final publications resulting from the research.

4.3 PROTAGE[22] (Preservation Organizations Using Tools in AGent Environments)

PROTAGE is addressing the challenges related to the preservation of digital resources of increasing volume and heterogeneity by developing tools allowing for more efficiency and self-reliance of preservation processes. For this purpose, PROTAGE researchers are exploring the value of a promising technology - software agents - for the automation of digital preservation processes. Based on the latest research on digital preservation strategies and on autonomous systems, the project intends to build and validate flexible and extensible software agents for long-term digital preservation and access that can cooperate with and be integrated in existing and new preservation systems to support various aspects of the digital preservation workflow such as the submission / ingestion of digital material, monitoring of preservation systems and transfer between repositories.

Tools developed by the PROTAGE project will:

- enable content producers to create and publish in a preservation-compatible manner,
- provide digital repositories with means of further automating the preservation processes,
- facilitate seamless interoperation between content providers, libraries and archives, and end-users throughout Europe.
- Targeted end-users are curators and digital content creators, including individuals managing their own digital collections. PROTAGE will use archive

and library materials from the project partners for system and user tests and external stakeholders in further validation.

4.4 PrestoPRIME[23]

PrestoPRIME will address long-term preservation of and access to digital audio-visual content by integrating media archives with European on-line digital libraries. Research will result in a range of tools and services, delivered through a networked Competence Centre.

4.5 SHAMAN[24] (Sustaining Heritage Access through Multivalent ArchiviNg)

The goals of the SHAMAN project are:

- Establish an Open Distributed Resource Management Infrastructure Framework enabling Grid-Based Resource Integration, reflecting, refining and extending the OAIS model and taking advantage of the latest state of the art in virtualisation and distribution technologies from the fields of GRID computing, Federated Digital Libraries, and Persistent Archives.
- Develop and integrate technologies to support Contextual And Multivalent Archival And Preservation Processes which are adapted and significantly extended from the fields of content and document Management and Information Systems.
- Develop and integrate technologies to support Semantic Constraint-Based Collection Management to target one of the key challenges in automating one class of digital preservation core functions.
- Support the Managing Of Future Requirements by securing Interoperability With Future Environments and Maintaining Essential Properties of the preserved content.
- Foster take up of DP technologies by facilitating suites of products and services able to attack problems and needs from an integrated point of view.

(Source: SHAMAN, Project website)

5 Relevant Framework 6 Projects

5.1 CASPAR[25] (Cultural, Artistic and Scientific knowledge Preservation, for Access and Retrieval)

- System framework for longer term preservation of heterogeneous digital resources.
- Foundation methodology for digital preservation activities
- Open Archival Information Systems Reference Model(OAIS, ISO 14721)
- Characterisation(representation and preservation description information), virtual storage, access services (intuitive query and browsing mechanisms)
- Standardisation, authentication, accreditation and digital rights management
- Test-beds: science, performing arts and tangible cultural heritage
- Participants: large research organisations, universities, audio-visual content, cultural heritage sites, technology companies

5.2 Digital Preservation Europe[26]

- Raise the profile of digital preservation (organisations, specialist professionals, citizens)
- Promote the ability of Member States acting together
- Use cross-sectoral cooperation to avoid redundancy and duplication of effort
- Ensure auditable and certificated standards for digital preservation processes are selected and introduced
- Facilitate skills development through training packages
- Enable relevant research coordination and exchange
- Develop and promote a research agenda roadmap
- Participants: organisations involved in ongoing national initiatives (e.g. DCC, DPC, nestor

5.3 P2P-FUSION[27] - Peer to Peer Fusion

P2P-FUSION addresses the current difficulties in the legal creative reuse of audio and video media in the internet environment. The project aims to create an open, accessible, legal and economically efficient environment for creative audiovisual cultural activities that can also be based on the work of others and on materials from cultural institutions, through built-in, easy-to-use support for suitable licensing schemes.

5.4 PLANETS[28] - Preservation and Long-term Access to our Cultural and Scientific Heritage.

PLANETS will deliver a distributed service environment for the management of digital information preservation, with a special focus on the needs of organizations having the preservation of social and cultural memory as a core task. This includes planning services, methodologies, tools and services for characterization of digital objects,

innovative solutions for preservation actions, and an interoperability framework to seamlessly integrate tools and services in a distributed service network. The project will enable organizations to improve decision-making about long term preservation, ensure long-term access to their valued digital content and control the costs of preservation through increased automation and a scalable infrastructure.

5.5 PRESTOSPACE[29] - Preservation towards storage and access.

Standardized practices for audio-visual contents in Europe Institutions traditionally responsible for preserving audio-visual collections (broadcasters, research institutions, libraries, museums, etc.) face major technical, organizational, resource related, and legal challenges in taking on the migration to digital formats and the preservation of already digitized holdings. The project has developed technical tools and a semi-automated integrated system, the 'Preservation factory', for digitization and preservation of all types of audio-visual collections.

6 Other Relevant Projects and Developments

6.1 Europeana[30]

Europeana is run by a core team based in KB. It builds on the project management and technical expertise developed by The European Library, which is a service of the Conference of European National Librarians.

The Europeana prototype is the result of a 2-year project that began in July 2007. Europeana.eu went live on 20 November 2008, launched by Viviane Reding, European Commissioner for Information Society and Media.

Europeana is a Thematic Network funded by the European Commission under the eContentplus programme, as part of the i2010 policy. Originally known as the European digital library network – EDLnet – it is a partnership of 100 representatives of heritage and knowledge organisations and IT experts from throughout Europe. They contribute to the Work Packages that are solving the technical and usability issues.

6.2 DPFuse (Digital Preservation Functions and Services)

From December 2007 until March 2008, a working group of Deutsche Nationalbibliothek (DNB), the Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) and the National Library of the Netherlands (KB) discussed the functional requirements for a next generation long-term preservation solution. The intention is to produce a statement of common requirements by 2012 (KB, DNB, & SUB, 2008)

6.3 The European Library (TEL)

TEL (Dobratz & Neuroth, 2004) is a collaboration between the national libraries of Finland, Germany, Italy, The Netherlands, Portugal, Slovenia, Switzerland and the United Kingdom, together with the Italian Central Cataloguing Institute (ICCU) It aims to establish a professionally designed and maintained single access point to selected holdings spanning a range of collections in all the partner national libraries so that the informed citizen in any country can utilise the resources not only of his or her own national library but also—during the same search session—the resources of any other partner national libraries that may hold material relevant to the citizen's interest.

6.4 nestor[31] (Network of Expertise in Long-Term Storage of Digital Resources)

nestor is a sub-project of "Competence Network New Services, Standardisation, Metadata (Kompetenznetzwerk Neue Dienste, Standardisierung, Metadaten)", funded by the German Federal Ministry of Education and Research (Deutsche Nationalbibliothek, 2009).

The DNB leads nestor, working in cooperation with the Bayerische Staatsbibliothek

München, Computer- und Medienservice / Universitätsbibliothek der Humboldt-Universität zu Berlin, the Fernuniversität Hagen, the Generaldirektion der Staatlichen Archive Bayerns, the Institut für Museumskunde, the Niedersächsische Staats- und Universitätsbibliothek Göttingen and the Bundesarchiv (Deutsche Nationalbibliothek, 2009).

The goals of nestor (Dobratz & Neuroth, 2004) are to:

- Increasing awareness. Awareness of the problems related to, and the exigency of, digital long-term preservation within museums, archives and libraries will be increased and this information will be disseminated to policymakers to convince them of the need for action in this field.
- Creating a network of information. Information concerning technical, organizational, legal and other aspects of digital long-term preservation will be acquired, and the existing information about current research, projects and "best practice"-results will be bundled and disseminated.
- Cooperating with national and international institutions. National (German) and international strategic alliances will be established between archiving institutions (libraries, archives, museums) on one side and industry and research facilities on the other in order to solve together mutual challenges with respect to the long-term preservation of digital resources.
- Coordinating strategies. Subject to the international context, an interdisciplinary forum will be set up for the development and coordination of strategies for long-term preservation of digital sources in the German cultural sector. Services, technologies and standards for long-term preservation of digital resources will be developed.
- Presenting a long-lasting organizational model. At the end of the project, nestor partners will present a concept for a long-lasting organizational model for the network of excellence. In addition, network partners will vote on which tasks are going to be shared and which will be completed solely by individual organizations, in view of the delimitation between libraries, archives and museums. In order to record the current state of long-term digital preservation efforts and current demands being made on the network of excellence, 7 expert reports were commissioned at the end of 2003. The first expert reports are expected to be completed by mid-2004.

6.5 The PRONOM Technical Registry[32]

This provides very detailed information and associated metadata for a wide range of file types as part of an active preservation strategy. Although more geared at present to migration, it does cater for file formats such as .exe files; provides a field for technical environment that aggregates software and hardware information (mostly blank presently); and creates an emulation pathway. Significantly, unique information is held with an object, so metadata can be determined once and for all, and does not need to be created afresh for every ingest. Container formats such as .zip and .tar files are currently accommodated by PRONOM, which iteratively extracts and processes them. Such an iterative approach is also employed to tackle complex items such as websites, but it is not always possible to capture all the associated semantics of a webpage. An inbuilt risk strategy highlights the extent to which formats are currently supported. Planned developments of PRONOM include

its extension to cover complex media such as CD-ROMs, for example. PRONOM uses an XIP metadata structure and metadata can be exported in a PREMIS format. Thus, the inclusion in any future emulation system of characterization software such as PRONOM / UDFR / GDFR could contribute greatly in terms of creating technical metadata and selecting suitable emulation pathways / tools / modules etc.

However, there are many challenges to consider:

- filling in the metadata fields would be very resource-consuming (with information sent into the scheme by outside users);
- common formats would need to be agreed upon;
- libraries may not want to rely upon an external registry, and may want their own version of it;
- a file extension does not always sufficiently determine a digital object, which may have one of several possible formats;
- expressions are important for libraries, in particular representation information needs to be understandable for future and as yet unknown designated communities;
- the original technical environment may be unknown;
- important characteristics captured now for migration may be different for emulation;
- there may be several emulation paths per digital object;
- re-characterization of objects may be useful where there are new interfaces, for example an emulation interface without a joystick;
- a metadata schema may become unreadable in ten year's time;
- characterization data can be larger than the original data file;

6.6 The Global Digital Format Registry (GDFR)[33]

The GDFR was announced in a 2006 press release from Harvard University as a new project supported by the Andrew W. Mellon Foundation "for the development of a registry of authoritative information about digital formats" (President and Fellows of Harvard College, 2006). According to Stephen Abrams[34], GDFR was planned to be "a sustainable service available to any preservation institution that chooses to participate. From the outset, we've envisioned the registry as a distributed network of individual 'nodes.' Each node will have a full copy of all the format-typing data in the GDFR. Carefully vetted information and updates will be distributed among the nodes following appropriate technical review. GDFR will also provide a separate track for distributing non-vetted information, so that problems and issues identified in the course of daily work can be quickly shared by participants." (Cited in: President and Fellows of Harvard College, 2006)

Some progress was made on GDFR, but as of Spring 2009, work had not begun database building (*Unified Digital Formats Registry (UDFR): Proposal and roadmap*, 2009).

6.7 The Unified Digital Format Registry (UDFR)[35]

The UDFR is an initiative announced in April 2009 which aims to build a single shared formats registry. The UDFR is to be based on the existing PRONOM system and database and will incorporate the GDFR.

6.8 IBM Preservation Manager[36]

In 2003 the KB began developing a preservation system for the e-Depot consisting of tools for permanent access, a Preservation Processor and a Preservation Manager. The Preservation Manager describes the technical environments required to support access to digital objects stored in the e-Depot. was to provide technical metadata. The Preservation Manager makes use of two key concepts:

- Preservation Layer Models (PLM)
  This describes the different layers of technology required to access a digital object and typically might consist of:
  - a data format layer
  - a viewer application layer
  - an operating system layer
  - reference platform layer

- View Paths
  This describes the PLM required for a specific object format.

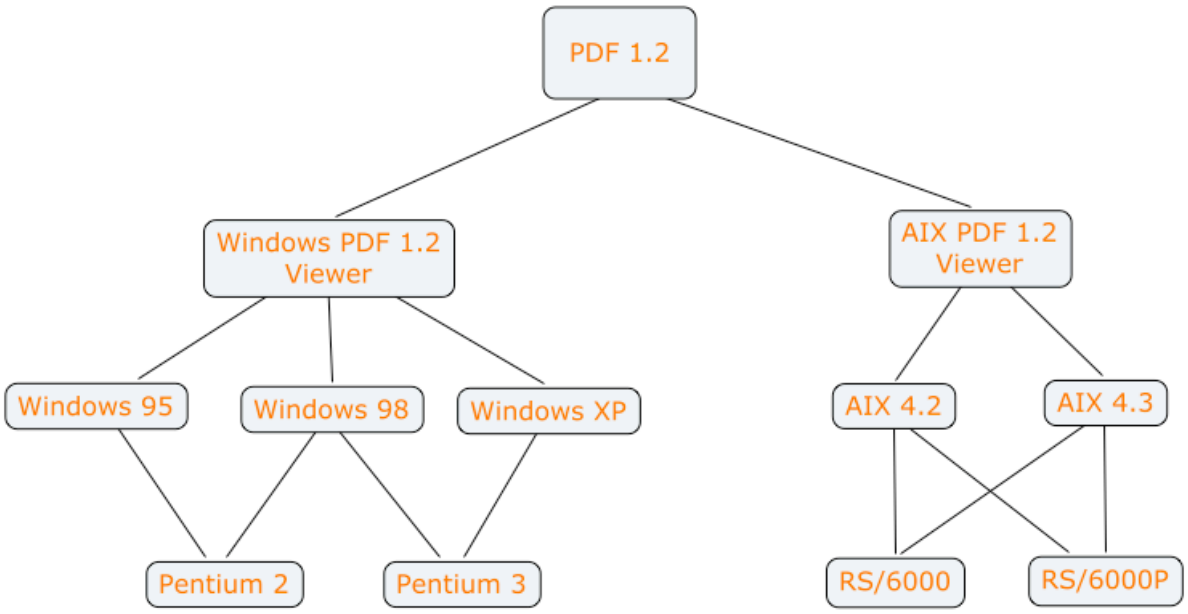

Fig. 14. Example View Path

(Based on: Brown, 2008)

“The work of the KB has several unique elements. Firstly, although few would dispute that emulation tools should be considered as elements of representation information networks, the Preservation Manager is probably the only RIR to address the modelling of emulation-based networks in any depth. Secondly, the KB plan to use analysis of view paths as the basis for assessing risks to digital objects. This risk assessment is based on the number of view paths currently available, together with other preservation variables, and is conceptually comparable with the PRONOM risk assessment service.” (Brown, 2008)

7 Institutional Issues

What is the state of the art regarding emulation in memory institutions? What also of their future policies: do they intend to move toward emulation, and if so, what initiatives do they have to develop technical metadata to fit in with their plans? The starting point for this report was to survey four of the KEEP partners: the Dutch, French and German national libraries and the German Computer Game Museum. Note that KEEP does not have an archive as a partner: however the SME partner Tessella[37] works closely with major national archives. We will briefly describe each institution and their main relevant collaboration / initiatives; their collections policies and the resulting range of data objects; and lastly their current systems and future plans, highlighting the part played by emulation and environment metadata.

7.1 Brief Institutional Overview

The BnF[38], the DNB[39], the KB[40] and the CSM[41] represent quite a range in terms of the issues studied, particularly in terms of legal deposit. All three national libraries are involved in digital preservation research and development at an international level, and contribute to the development of international library standards. The KB has a specific focus on supporting the Dutch national scientific information infrastructure[42]; the DNB is the leading partner in developing and maintaining bibliographic rules and standards in Germany[43]; and the BnF is collaborating on bibliographic databases[44].

7.2 Current European Research Projects / Collaboration

The KB and the BnF are active partners in Europeana[45] which provides access to pan-European digital cultural heritage content; and IMPACT, which seeks to improve access to text via advancements in OCR techniques. The DNB are partners in nestor[46]. The DNB and KB participate in PARSE.Insight, the KB, Tessella are partners in Planets[47], which is looking at, for example, emulation of hardware, on top of which migration can be carried out. The KB participates in IPIL (Integrating Planets into Libraries). The DNB has a major role in SHAMAN. The DNB and KB collaborate on DPFuse: a cooperation of 8 European national libraries to set up requirements for a “next generation” LTP system to be implemented in 2012. The KB is a member of DRIVER and DRIVER2, which look into interoperability of digital (enhanced) publications between digital repositories in Europe, and also takes part in Crossref, a cooperation between publishers and preservation services to open up titles in case of trigger events.

7.3 Collection Policies

There is considerable diversity in terms of what each institution collects, and this has a direct bearing on their respective digital preservation strategies.

The institution with the broadest array of digital objects is the BnF, which is a legal depot since 1938 for phonograms; 1975 for video and multimedia; 1992 for audio visual and electronic documents; and it has been a web legal deposit since 2006. They hold a wide range of multimedia, from the 1960s (book plus tape and disk) to the present day. In particular, the BnF keeps computer games, a key factor

regarding their interest in emulation.

The DNB is responsible for keeping publications produced in Germany since 1913; German- language publications issued abroad since 1913; translations of German works, foreign-language works about Germany published abroad since 1913, to administer the Deutsches Exilarchiv 1933-45, the Anne-Frank-Shoah-Bibliothek and the Deutsches Buch- und Schriftmuseum ("Gesetz uber Die Deutsche Nationalbibliothek," 2006 §2.1 & §2.2).

Since 2006, the DNB have also been required to keep online (web) publications ("Gesetz uber Die Deutsche Nationalbibliothek," 2006 §3.1 - §3.3). Prior to 2006, the DNB had been collecting networked digital publications on a voluntary basis.[48] The DNB use automated methods to capture much of the German part of the web. However they do not collect computer games. By contrast, the CSM does hold computer games (Floppy Disks, video, CD-ROM, DVD, cartridge…), computer hardware, magazines, and video.

The KB preserves publications with Dutch imprint since 1974; scientific publications (e-journals and e-books); digitized material; websites; and multimedia applications (CDROMs). They do not hold utility software, computer games, operating systems or web browsers.

7.4 Current / Future Preservation Systems at the BnF

The BnF comprises a research library and a public library, each posing different technical problems that need to be addressed separately. Currently the research library has a near online system, where the user asks online for a document, and the general service department (back office) responds to their request via a secure catalogue system. In terms of organization, the BnF has three main sections: Collections; Networks and Services; and Administration and Human Resources. The main focus of this survey is the Audio Visual (AV) department, which is part of Collections and has more specific computing needs for digital preservation than other departments in that section. In 2001 the AV department commenced the digitization of all their audio material, starting with video tapes. However, granting access to the digital version of the whole video collection would not appear to be a practical proposition currently. Audio CD then DVD are next in line for digital transfer. In all, 80 TB of data will need to be migrated to the BnF's new system SPAR (see below).

For the AV department, there is no formal metadata processing beyond file-type recognition, comprising characterisation via a catalogue link of a file as .wav or .mpeg file, for example. Each packet or document in the collection has information on the tape or card used, but this is not formal technical metadata as such, and is only used if a problem in the series is met. For .mpeg2 files, the file header gives all the necessary information. For audio and video material this is quite basic, but in terms of metadata, a DVD is a complex, variable digital object, whose structure is not currently known. Secure access to and reliable definition of such multimedia is part of the raison d'être for the BnF's involvement in KEEP. Currently the AV department has an interesting and valuable system[49] which uses 3rd party file transfer tools[50], and available in the reading room of tower 3, whereby users may access a limited set of digital objects running under emulation. The process is partially automated and requires an emulator suitable for the selected digital object to be selected by an administrator before the digital object is made available to the user. There is a database behind this process containing some metadata.

A new long-term preservation system is currently under development at the BnF which encompasses both migration and emulation. SPAR (Distributed Archiving and Preservation System) (http://bibnum.bnf.fr/spar) is directed by the Information Systems Department, part of the Networks and Services section, and the first phase of SPAR development has ended with the hardware now in place, and the software almost ready. SPAR is a modular system comprising both open source and proprietary software, and using RDF (Resource Definition Framework): grid technology. SPAR will include web archiving, and plans to use Java for its web services, with a link to the Gallica catalogue. The intention is to integrate all AV department material into SPAR, which presents some challenges as the AV department has more of an emulation-only standpoint: seeking to save, not change the original bitstream (Excel, ppt etc.), whereas the SPAR long-term plan tends currently to focus more on migration, which transforms the bitstream.

SPAR will be OAIS-compliant (OAIS standard ISO 14721) with rights management. METS and PREMIS-compliant metadata will be built for SPAR to comply with a detailed metadata specification. SPAR will use Virtuoso software for metadata management. In particular, the AV data needs to be integrated into METS. Specifically, the METS packaging will comprise both strongly and weakly-linked metadata; descriptive metadata will be qualified Dublin Core; preservation metadata will be PREMIS and rights information will be ODRL (Open Digital Rights Language). Technical metadata will depend on the channel: MIX for still images, textMD for text being possible candidates. Four levels of granularity are envisaged: a set which is an intellectual grouping (collection, periodical, …); a group which is a digital object (monograph, series of images, …); an object which is a digital element (a page, an image, a track, …); and a file which is a data-object (digital file or bit-stream). An object will have two representations: a literal description comprising text introducing a tool or process; and a formal description made up of an XML file based on the Environment PREMIS tag, the exact version of the tool, and the execution platform. In addition there will be a data model to map from METS to RDF.

7.5 Current / Future Preservation Systems at the KB

The KB e-Depot system (http://www.kb.nl/dnp/e-depot/e-depot-en.html) uses the IBM DIAS system as its core. With a capacity of 8,000 titles, 13 million articles, 10 TB data, the DIAS system is able to load 10,000-40,000 publications per day, with an annual growth of 1.5-2 million e-journal articles. Material currently processed includes e-journals, dissertations and e-books. Data is delivered via zip files, and is pre-processed into a standardized SIP file. The e-Depot system is based on the OAIS-reference model (ISO 14721:2003), and uses extended Dublin core bibliographic metadata (15/16 fields), which is converted to internal KB format using stylesheets; XSLT being used to convert the XML to the KB internal metadata standard. Bibliographic metadata is stored in OCLC/PICA format in the KB catalogue. Whilst digital object bibliographic metadata can be obtained automatically, describing the environment currently poses a difficult challenge, with the result that technical metadata is created manually. Web archiving is in progress, although currently it is proving difficult to obtain reliable technical metadata for web data.

The e-Depot has a specific workflow. The publishers send samples based on a checklist and PDF guidelines. The metadata organization is then analyzed (file formats, batch structure), and the pre-process is then adjusted, with conversions performed according to the e-Depot DTD (Data Transfer Description). Next testing is carried out, followed by the signing of the archiving agreement. Data is then sent by the publishers via FTP, tape or CD-ROM, and the corresponding titles and ISSN

numbers are entered into the catalogue. The ingest procedure is then started by the creation of a SIP. This system works well on the whole, but it is difficult for publishers to adjust their whole workflow to accommodate any metadata in the wrong format, leaving this to the metadata organization and pre-process functions.

Currently the KB is at the forefront of research into incorporating emulation into its digital preservation system, as can be seen with the e-Depot Preservation Manager. (Brown, 2008) states that "although few would dispute that emulation tools should be considered as elements of representation information networks, the Preservation Manager is probably the only RIR (Representation Information Registry) to address the modelling of emulation-based networks in any depth". (Hoorens, Rothenberg, van Oranje, van der Mandele, & Levitt, 2007, p. 63) reports that the "KB's unique, multi-pronged technical strategy for preservation includes migration, the use of IBM's Universal Virtual Computer (UVC) to perform 'data preservation' and the use of emulation." Spearheading the KB's emulation initiative is the KB / Dutch National Archive joint Dioscuri project[51] comprising an emulator specifically designed for digital preservation (van der Hoeven, Lohman, & Verdegem, 2007), whose development is continuing under the auspices of the Planets project.

In terms of a future system, the KB plans to extend e-Depot, as it will need a new system and new data model following the results of the DPFuse project. New content types will include complex, compound and enhanced digital objects, such as Word or Excel files; dynamic digital objects; and MP3, jpeg; Tiff files, etc. Data currently held on CD-ROMS, cartridges, and floppy disks is very diverse and ranges from Wordperfect 5.1 documents to multimedia applications for Windows environments. Websites will also be stored in the e-Depot and these can contain any kind of object. Similarly for enriched publications such as PDFs with embedded scripts and movies. An extra difficulty to be faced in the future is that most data carriers are stored on the shelves together with the physical publication with very little technical metadata. Rapid growth in the quantity of digital data stored is anticipated, reaching around 500 TB in 2012. An extended e-Depot would need to check object type using characterisation software such as PRONOM, DROID (Digital Record Object IDentification) and JHOVE (JSTOR/Harvard Object Validation Environment). Two copies of metadata will be produced for the extended e-Depot: one standardized to the KB catalogue, and one in the original form stored with the digital object.

7.6 Current / Future Preservation Systems at the DNB

The DNB and KB were both identified in a 2006 survey as being two out of only three organisations worldwide that were then actively researching emulation as an archiving strategy (Kenney, Entlich, Hirtle, McGovern, & Buckley, 2006) The DNB is currently using the kopal system[52], a German project developed from 2004 to 2007 in collaboration with the Goettingen State and University Library, GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Goettingen) and IBM. The kopal system contains of the OAIS compliant[53] kopal-DIAS[54] (Digital Information Archiving System) and the open source Java tools koLibRI[55]. Within the koLibRI umbrella, metadata generation is carried out with JHOVE (JournalSTORage/Harvard Object Validation Environment)[56] software. The SIP and DIP data packages are transferred over the SIP and DIP interfaces to DIAS. The Interface Items conform to the Universal Object Format[57] for kopal-DIAS.

In addition to the Universal Object Format, the DNB employs URNs[58] (Uniform Resource Names), which guarantees that a resource has a persistent unique identifier, with at least one functioning URL assigned to each URN. The DNB also

uses DAFFODIL[59], a virtual digital library system to aid user information search and retrieval. In particular it facilitates a range of search functions using diverse metadata annotations in multifarious digital libraries.

The DNB currently employs a very wide range of metadata standards, starting with Dublin Core for bibliographic data. MetaDiss[60] is used for theses and online dissertations, whilst XMetaDiss[61] is a format coordinated under the auspices of German and other university libraries that has the particular strengths of providing a hierarchical structure, interoperability with international metadata search engines, and data exchange via OAI protocol. There is also the potential to use XSLT to transform XMetaDiss into other metadata formats. XMetaPers, the part of XMetaDiss related to persons, is also used. LMER[62] (Long-term preservation Metadata for Electronic Resources) is of particular importance here as it is technical metadata based on the National Library of New Zealand model, specifically developed by the DNB for the long term (see above). METS[63] is used for encoding descriptive, structural and administrative metadata. The German standardMAB2[64] is a format allowing automatic exchange of bibliographic metadata between libraries, and the associated MABxml[65] is also used. The international standard MARC21[66] (and MARCxml[67]) will replace MAB in the next few years, whilst ONIX[68] is used for representing information on books. In terms of acquiring metadata, automatic harvesting is easier from METS applications with classical indexing employing linguistic and statistical methods. Automatic metadata harvesting will be much more difficult in the future where the provenance and file formats of digital objects may not be well known.

The current workflow for printed material involves librarians manually recording information on each piece of printed material in the DNB collection. This is accomplished by a multi stage process with basic information (title, author, etc.,) being entering by relatively junior staff with more complex classifications being determined and recorded by more experienced and senior colleagues. A similar process is needed for digital material, but currently there is no agreed workflow for recording specific technical metadata about CD-ROMs, for example. This is a major drawback and it would be useful if one of the outputs of the KEEP project were such a workflow. An important practical point is that librarians who catalogue CD-ROMs do not run the software contained on them, and this can give rise to cataloguing problems such as a number of CD-ROMs designed for the Sony ebook being miscataloged as being for the IBM PC. In particular, librarians would welcome specific guidance regarding technical requirement priorities: for example preferred / minimum / acceptable versions of software or hardware. There is no emulation system per se currently at the DNB, but just a few legacy emulation programs bundled with old floppy disc images on CD-ROMs.

The present system contains approximately 85 TB and is shared with Goettingen State and University Library. A follow-up project will have ca.100 TB with the potential for ca.600 TB, to house 75, 000 online dissertations; 454 e-journals; over 2,000 monographs from Springer (Heidelberg, Berlin); 60 newsletters; 4,500 electronic publications from around 210 commercial and non-commercial publishers; web sites; the music library containing digital music files; CDs; sheet music (not yet digitized, potentially ~450 TB); and multimedia CD ROMs (potentially ~60 TB). Specifically there is a need for a reliable procedure to make CD-ROM images, as CDs are fragile with approximately 10% of the present music collection being unreadable. A future system would be METS and PREMIS compliant, and it is seen as vital that future metadata schemes are left sufficiently flexible to cope with any possible formats that the DNB may be responsible for collecting. A clear, metadata-driven scheme is seen as being desirable, that is both wide-ranging and easy to use,

with a very flexible XML file format and the ability to query every field. Both emulation (and the corresponding environment metadata) and migration are seen as a vital part of this future development, with the proviso that the DNB will be looking closely at the results of the KEEP legal study before deciding how to proceed on emulation.

The DNB's involvement in the SHAMAN project, with its input from the scientific and engineering communities, has brought to light some interesting issues that are salient here. First, the Grid-based production system: is this a short, medium or long-term trend? Next, the need for interfaces to both SHAMAN and legacy systems is key, and kopal is the main test case for integration, providing the test collections. A too-narrow focus on the core OAIS strength is perceived as a current limitation, and there is a lot of work currently on digital preservation where digital mechanisms lie outside OAIS, for example harvesting metadata before ingest, which would require publisher cooperation. This calls for a broadening of the view (see fig. below) to the left and right of OAIS, the impact of whose second version, recently released, is of vital importance to all institutions in this survey. Finally, metadata was considered at the beginning of the SHAMAN project, when the relevant processes were identified and clustered into phases of the digital object; interfaces were then defined; and exchanges developed. It was seen as vital here that consideration of metadata did not constrain development, and the abstract data models behind serialisation and metadata were both taken into account.

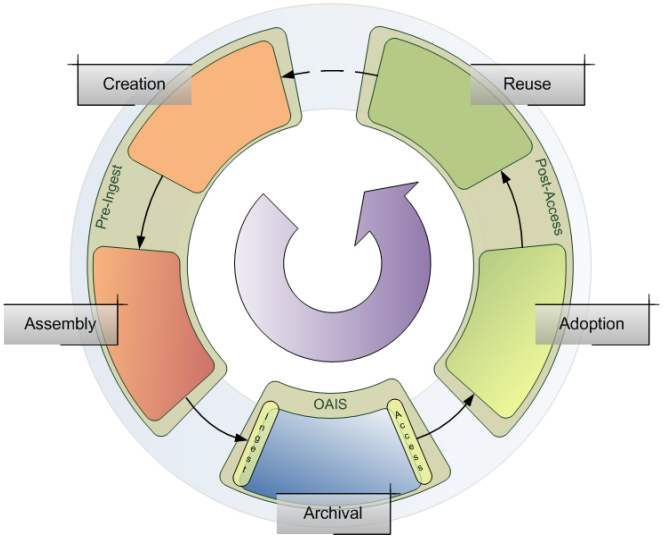

Fig. 15 A dimension of the new Reference Architecture for SHAMAN (Source: SHAMAN, Project website)

Lastly, developments in web archiving are also pertinent, as they involve the analysis of complex and dynamic digital objects whose technical environment is challenging to describe. Some interesting generic metadata definitions have emerged from netpreserve.org, archive.org and the WARC (Web ARCive file format)[69] standards. URNs also play an important role in web-archiving, and an Australian digital preservation working group is currently looking at browser emulation[70].

7.7 Current / Future Preservation Systems at the CSM

The CSM is currently in collaboration with DIGAREC (Digital Games Research Centre at the University of Potsdam), and is an external partner to the Planets project.

The current system comprises a web server onto which metadata is entered manually.

The software metadata includes:

- Title
- Original Title
- System; Language
- Data Medium
- Publisher; Developer
- Genre; Subgenre
- Product Type
- Release Age
- Archive number
- EAN (European Article Number)
- First Published
- Country
- Manual
- Remarks

The metadata for the hardware is stored in a separate database, which is only accessible offline. The CSM also uses a proprietary museum magazine database. MESS, MAME and others emulators are currently used for digital games emulation. A future system will have a common database for software and hardware, which will be accessible over the Internet. Following the instigation of KEEP, the new system will house the image files of transferred software.

In terms of current practice, there is a central computer games metadata registry database[71], which is connected to Media Control GfK International[72], and to which publishers send their metadata information at no cost. It is possible that this registry system may have synergies with other registry systems such as UDFR / GDFR which could useful in determining a metadata standard for complex multimedia objects such as CD-ROM. Each computer game has an EAN code, with different games releases having different EANs. Games released in Germany are assigned an EAN number and are then sent to USK (UnterhaltungssoftwareSelbstkontrolle – Maintaining software self control) data bank, where they get an H-rating (linked to zavatar). Such discussion brings to the fore the whole discussion regarding the overlap of technical registries and metadata creation.

8 Preservation Metadata For Games

> "Computer games, like other digital media, are extremely vulnerable to long-term loss, yet little work has been done to preserve them. As a result we are experiencing large-scale loss of the early years of gaming history. Computer games are an important part of modern popular culture, and yet are afforded little of the respect bestowed upon established media such as books, film, television and music. We must understand the reasons for the current lack of computer game preservation in order to devise strategies for the future. Computer game history is a difficult area to work in, because it is impossible to know what has been lost already, and early records are often incomplete."
>
> (Gooding & Terras, 2008)

8.1 Recent research in Germany

One of the most comprehensive studies so far undertaken on the practical problems involved in archiving computer software is to be found in Karsten Huth's Master's degree thesis (Huth, 2004b) Although Huth restricts himself to looking at the Atari Video games console (VCS 2600) which was released in the late 1970s[73] and the Commodore C64 which was first released in late 1981[74], his work is of continuing relevance to KEEP. It should be noted at the outset that the hardware platforms which Huth considers are relatively straightforward and stable compared to, for example, IBM PC (clones) of the same period and consequently present a more tractable problem than are presented in archiving software for more complicated and dynamic hardware platforms. That said, much of the information which Huth advocates capturing is typical to software generally.

Huth's proposed metadata set for the AIP runs to over 100 elements and includes:

- Descriptive information for storage and retrieval
- Archival metadata about ingest, storage, use and fault reporting
- Instructional metadata for users to run the game
- Technical requirements and dependencies, including forward-facing dependencies for product extensions, fixes, alterations and expansion packs.

His approach involves combining features of several existing schemes: OCLC (*Digital Archive Metadata Elements*, 2003), Dublin Core (DCMI, 2003), MAD, (Procter & Cook, 2000), DIN (German Institution of Standards) 66230 and DiGA (DiGA, 2003), together with some extra elements of his own.

Conceptually, the model breaks down into three sections. First, there is descriptive metadata, of use to both administrators and end-users. Secondly, there are numerous fields for administration; ingest and storage information plus report fields for debugging and faults. These comprise a large portion of the model. Finally, technical information covers hardware components, software requirements and dependencies. Huth's model allows for bidirectional requirements for add-ons, patches and extensions to the original software. The technical information is broken down according to ingest, storage and run-time requirements. Instructions for users to both install and run games are included as separate metadata fields that cover both technical requirements and the practical business of running the media.

Emulation is referenced directly by Huth (Huth, 2004a, 2004b) and the metadata system proposed does seem to include most technical requirements necessary to

enable emulation to occur within the hardware domains he considers Additionally, the archival metadata structure is extensive and suitable for archiving. Huth's study predates PREMIS but does make use of elements of the CEDARS model.

A more detailed study with full German-English translation is necessary to fully check compatibility with PREMIS, although it seems likely that, given the complexity of the model, there will be a high degree of commonality in most descriptors. Overall, it represents the most comprehensive metadata model for games currently available, and adaptability is driven primarily by the issue of ensuring easy adoption by archives as a supplement to current procedure.

9 Non-Preservation Metadata For Games

9.1 MobyGames[75]

MobyGames is a large and well-used games database providing descriptive information about current and obsolete games alongside professional and community reviews, screenshots, developer information and some technical specifications. Data is user-generated within a metadata template. This is focused purely on descriptive data and is far simpler than the preservation models described above.

Each object requires the following descriptive data:

- Title
- Release date
- Release country
- Developer
- Publisher
- Platform
- Genre
- Entry data

Optional data includes:

- Additional Release Dates and Countries
- PEGI/ESRB Rating
- Developer Credits (including, for example, Porting, Sound, Graphics, Localisation, Engine and Middleware developer)
- System Requirements
- Community Ranking
- Professional Critics' Ranking
- Number Of Players Supported
- Supporting Forums
- Boxshots
- Screenshots
- Trivia Pages
- Additional Documents (such as manuals)
- Links to External Purchasing and Download Sites.

The required data is essential for descriptive metadata for objects in the KEEP system, as are the System Requirements, although these may be more sporadic as they are only optional fields. One key omission in descriptive data is copyright information. The credits, if they are to be included within KEEP, will necessitate an extension of the AUTHOR and PUBLISHER fields. In addition to this, middleware components are not listed, which may also require a new field and potential sub-fields for further accreditation. The GENRE field is pre-determined and includes 8 basic genre classifications:

- Action
- Adventure
- Educational
- Racing / Driving
- Role-Playing (RPG)

- Simulation
- Sports
    - Baseball
    - Basketball
    - Bike/Bicycling
    - Bowling
    - Boxing
    - Cricket
    - Fishing
    - Football (American)
    - Golf
    - Hockey
    - Horse / Derby
    - Hunting
    - Motorcycle
    - Off-Road / Monster Truck
    - Olympiad
    - Paintball
    - Ping Pong/Table Tennis
    - Pool / Snooker
    - Rugby
    - Sailing / Boating
    - Skateboarding
    - Snowboarding / Skiing
    - Soccer / Football (European)
    - Surfing
    - Tennis
    - Tricks / Stunts
    - Volleyball
    - Wakeboarding
    - Wrestling
- Strategy

Plus additional classifications for :

- Non-Sports
- Adult
- Anime / Manga
- Arcade
- BattleMech
- Board / Party Game
- Cards
- Casino
- Chess
- Comics
- Cyberpunk / Dark Sci-Fi
- Detective / Mystery
- Fighting

- Flight
- Game Show
- Helicopter
- Historical Battle (specific/exact)
- Horror
- Interactive Fiction
- Interactive Fiction with Graphics
- Interactive Movie
- Managerial
- Martial Arts
- Medieval / Fantasy
- Meditative / Zen
- Mental training
- Naval
- Paddle / Pong
- Persistent Universe
- Pinball
- Post-Apocalyptic
- Puzzle-Solving
- Real-Time
- Rhythm / Music
- Sci-Fi / Futuristic
- Shooter
- Spy / Espionage
- Stealth
- Survival Horror
- Tank
- Train
- Turn-based
- VideBackdrop

- Rating Systems
  - 3DRating
  - BBFC Rating
  - CCE (IGAC) Rating
  - CERRating
  - DJCTQ Rating
  - Dejus Rating
  - ELSPA Rating
  - ESRB Rating
  - FPB Rating
  - IFCRating
  - KMRB Rating
  - Kijkwijzer Rating
  - OFLC (Australia) Rating
  - OFLC (New Zealand) Rating
  - PEGI Rating
  - RSAC Rating

    - SELL Rating
    - USK Rating
    - VET/SFB Rating
    - VRC Rating
    - aDeSe Rating
- Companies (366 in total)[76]

Other Attributes (6 fields, including Emulator, Editor and Add-On).

The genre classification is an important field, but the most problematic in terms of achieving a consistent and compatible system, and the approach taken by MobyGames appears generally rather organic.

The technical data is limited. It is possible to search the database according to platform and system requirements (such as minimum CPU requirement, minimum RAM and supported controllers) which narrows the field of search, but this is not a systematic list for the reasons outlined above. KEEP would require these technical requirements to be a required field of entry in order to serve the emulation framework.

9.2 Gamespot[77]

Gamespot is a commercial games information site that includes news and reviews, game information, patches and cheats (and other downloads), developer and publisher information, audiovisual supporting material, downloadable demos and supported community pages such as forums and blogs.

Data about each game is mainly descriptive, with the following fields:

- Title
- Developer
- Publisher
- Release Date
- Genre[78]
- PEGI/ESRB Rating
- Platforms.

Technical information is very limited. Search cannot be narrowed beyond platform, and the information held replicates basic information on the object itself[79]:

- Operating System
- Ram
- Cpu
- Hard Drive Space
- Video Memory
- Direct X Version

Overall, Gamespot is aimed at game users who either own the object already or are looking for supplementary information such as reviews, previews and fora surrounding it. Although it may be worth cross-referencing descriptive information to ensure accuracy, its use beyond this is limited. The one advantage is that, unlike MobyGames, the addition of objects to the database is commercialised and thus it is a better resource for information about recent releases, and the addition is more systematic.

10 Abandonware Sites

Abandonware refers to obsolete or historical software that still falls under copyright protection but is deemed to be no longer an object of interest to its developers or publishers. This distinguishes it from freeware or shareware, which is explicitly public domain. Khong defines abandonware as follows:

"Orphan works and abandonware can be generally defined as copyrighted works which are still within their terms of protection but are no longer commercially available to the public. If the copyright owner is available and willing to license the work, the work is not considered abandoned even though no commercial copies are for sale. On the other hand, if the copyright owner sets unreasonably onerous licensing terms in order to discourage the supply of his copyright work, the work may rightly be considered as abandoned" (Khong, 2007)

Abandonware sites currently supply free copies of objects over the Internet. Whilst this undoubtedly breaches copyright law, most operate on the implicit understanding that if objections arise, objects are withdrawn. As such, the abandonware community implicitly distinguishes itself from warez trading, or the free sharing of copyrighted material, which would normally fall under the category of piracy. However, according to Goldman, although "Some abandonware enthusiasts consider themselves historians or archivists, but in all other respects their actions are indistinguishable from other warez traders." (Goldman, 2004)

In real terms, although abandonware sites operate semi-legally at best, they nevertheless form a serious, non-commercial and non-institutional basis for preservation of games and have ensured that regardless of their legal status, copies of obsolete games that may otherwise have been lost, have been retained. Generally, metadata on games is restricted to descriptive data only, with a limited environment set that may link through to emulators or suggest run environments. The two sites detailed below are typical of most abandonware sites. They are included here to offer an illustration of how such sites are handling metadata and archiving, although it should be noted that the legal issues surrounding abandonware preclude any direct relationship with the KEEP project at this point.

10.1 Abandonia[80]

Abandonia is a community site dedicated to preserving access to obsolete games. This means it has some relationship to emulation, but suffers from the normal legal issues surrounding this type of database. Objects are uploaded by the community without any systematic preservationist agenda. As such, the metadata held on objects is simplistic and predominantly descriptive. Once again, the standard fields are included: :

- Title
- Producer (Developer)
- Publisher
- Year Of Release
- Size.

Abandonia concentrates on DOS gaming but provides links, where appropriate to other fan-preservation sites if alternative operating system versions of a title are available. Screenshots, descriptions and fora are included. Rather than genre, Abandonia uses a KEYWORDS field in which genre descriptions predominate but is

user-defined rather than systematically pre-defined. Finally, the site also contains a field suggesting multiple compatible emulators for the object in question.

Although the metadata present in this site is both highly limited and unsystematically defined, it nevertheless represents a stable and developed example of a community-driven games preservation site. In essence, it serves a similar public community to the extended user base of KEEP and the assumptions made about these users, their capabilities and their requirements may be of interest.

From a KEEP emulation metadata perspective, it may be a useful source of data to supplement or cross-check the information held by MobyGames and does, importantly, offer a link to non-IP protected objects for study and potential ingest.

10.2 Home of the Underdogs (HoTU)[81]

Currently undergoing redevelopment, HoTU, is similar in nature and features to Abandonia. Once again, a limited and unsystematic set of descriptive metadata is available to users:

- Title
- Year
- Developer
- Publisher
- Theme
- Multiplayer (Y/N)
- Designer
- System Requirements
- Copyright Information
- Where to Get It[82]

Some observations are in order here. The 'System Requirements' information which is held is extremely limited. However, two fields do stand out as interesting, first 'Copyright Information', which may be valuable for KEEP generally, and secondly 'Where to Get It' – an externally linking field directing the user to other community preservation sites to obtain actual media objects. This presents an interesting model for KEEP to consider, even if the metadata offered in the site of extremely limited.

11 Abbreviations

aDeSe	Asociación Española de Distribuidores y Editores de Software de Entretenimient[83]
AHDS	Arts and Humanities Data Service[84]
AIP	Archival Information Package[85]
APAIS	Australian Public Affairs Information Service
API	Application programming interface
ARKIS	The Archival Information System of the National and Regional State Archives
ASCII	American Standard Code for Information Interchange
AUDIOMD	Audio Metadata
Auslit	Australian Literature
AV	Audio Visual
BBFC	British Board of Film Classification
BnF	Bibliothèque nationale de France[86]
C64	Commodore 64[87]
CASPAR	Cultural, Artistic and Scientific knowledge Preservation, for Access and Retrieval[88]
CCE	COMISSÃO DE CLASSIFICAÇÃO DE ESPECTÁCULOS[89]
CCSDS	Consultative Committee for Space Data Systems[90]
CD-ROM	Compact Disc-Read-Only Memory[91]
CD-RW	Compact Disk - Re-Writable[92]
CEDARS	CURL Exemplars in Digital ARchiveS[93]
CLOCKSS	Controlled Lots of Copies Keep Stuff Safe
CPU	Central Processing Unit[94]
CROS	Cross Czech a.s.
CSM	Computerspiele Museum[95]
CURL	Consortium of University Research Libraries[96]
D.C	District of Columbia
DAITSS	Dark Archive In The Sunshine State[97]
DCC	Digital Curation Centre[98]
DCMI	Dublin Core Metadata Initiative[99]
DIAS	Digital Information Archiving System[100]
DiGA	The Digital Game Archive[101]
DIGAREC	Digital Games Research Centre[102]
DIN	Deutsches Institut für Normung[103]
DIP	Dissemination Information Package[104]
DJCTQ	Departamento de Justiça, Classificação, Títulos e Qualificação[105]

DLF	Digital Library Federation[106]
DNB	Deutsche Nationalbibliothek[107]
DOS	Disk Operating System[108]
DPC	Digital Preservation Coalition[109]
DPFuse	Digital Preservation Functions and Services
DRIVER	Digital Repository Infrastructure Vision for European Research[110]
DROID	Digital Record Object Identification
DSEP	Deposit System for Electronic Publications[111]
DTD	Document Type Definition[112]
DTD	Data Transfer Description
DVD	*Digital Versatile Disc* [113]
DVD-RW	Digital Versatile Disc Re-Writable[114]
e-book	Electronic Book[115]
EAN	European Article Number[116]
EDLnet	European digital library network[117]
EGDF	European Games Developer Federation[118]
eLib	Electronic Libraries
ELSPA	Entertainment & Leisure Software Publishers Association[119]
ESRB	Entertainment Software Rating Board[120]
FCLA	Florida Center for Library Automation[121]
Fedora	Flexible Extensible Digital Object Repository Architecture[122]
FORTRAN	FORmula TRANslation[123]
FPB	Film and Publication Board[124]
FTP	F*ile Transfer Protocol*[125]
GDFR	Global Digital Format Registry
GWDB	Gesellschaft für wissenschaftliche Datenverarbeitung mbH Goettingen[126]
HoTU	Home of the Underdogs[127]
IBM	International Business Machines[128]
ICCU	Italian Central Cataloguing Institute
IMPACT	Improving Acess to Text[129]
ISO	International Organization for Standardization[130]
ISSN	International Standard Serial Number[131]
JHOVE	JSTOR/Harvard Object Validation Environment[132]
JISC	Joint Information Systems Committee[133]
JOG	Joguin SAS[134]
JSTOR	Journal Storage[135]
KB	Koninklijke Bibliotheek[136]
KEEP	Keeping Emulation Environments Portable[137]

KMRB	Korea Media Rating Board[138]
koLibRI	kopal Library for Retrieval and Ingest[139]
kopal	Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen[140]
LiWA	Living Web Archives[141]
LMER	Long-term Preservation Metadata for Electronic Resources[142]
LOCKSS	Lots of Copies Keep Stuff Safe[143]
MAB	Maschinelle Austauschformat für Bibliotheken[144]
MAD	Manual of Archival Description[145]
MAME	Multiple Arcade Machine Emulator[146]
MARC	MAchine-Readable Cataloging[147]
MESS	Multiple Emulator Super System[148]
METS	Metadata Encoding and Transmission Standard [149]
MIX	Metadata for Images in XML [150]
MODS	Metadata Object Description Schema
mpeg	Moving Picture Experts Group[151]
MS	Microsoft[152]
NAS	National Archives of Scotland [153]
NBD	National Bibliographic Database[154]
NDNP	National Digital Newspaper Program[155]
NEDLIB	Networked European Deposit Library[156]
nestor	Network of Expertise in long-term STOrage and availability of digital Resources in Germany[157]
NISO	National Information Standards Organization[158]
NLA	National Library of Australia [159]
NLNZ	National Library of New Zealand[160]
NPO	National Preservation Office[161]
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting[162]
OAIS	Open Archival Information System[163]
OCLC	Online Computer Library Center[164]
OCR	Optical Character Recognition[165]
ODRL	Open Digital Rights Language[166]
OFLC	Office of Film and Literature Classification[167]
ONIX	Online Information Exchange[168]
OPAC	Online public access catalog[169]
OS	Operating System[170]
P2P	Peer-to-Peer[171]
PANDAS	PANDORA Digital Archiving System[172]

PANDORA	Preserving and Accessing Networked Documentary Resources of Australia[173]
PARSE	Permanent Access to the Records of Science in Europe[174]
PC	Personal Computer[175]
PDF	Portable Document Format[176]
PDI	Preservation Description Information[177]
PEGI	Pan European Game Information[178]
PID	Persistent Identifier[179]
PLANETS	Preservation and Long-term Access through Networked Services[180]
PM	Person Month[181]
ppt	Powerpoint[182]
PREMIS	PREservation Metadata Implementation Strategies[183]
PRESTOSPACE	Preservation towards storage and access: Standardised Practices for Audiovisual Contents in Europe[184]
PROTAGE	PReservation Organizations using Tools in AGent Environments[185]
PURL	Persistent Uniform Resource Locator[186]
QCL	Quantum Computation Language[187]
RAM	Random-Access Memory[188]
RDF	Resource Definition Framework[189]
RIR	Representation Information Registries[190]
RLG	Research Libraries Group[191]
RPG	Role Playing Game[192]
RSAC	Recreational Software Advisory Council[193]
Sci-Fi	Science Fiction[194]
SHAMAN	Sustaining Heritage Access through Multivalent ArchiviNg[195]
SIP	Submission Information Package[196]
SME	Small and medium enterprises[197]
SPAR	Système de Préservation et d'Archive Réparti[198]
SUB	Staats- und Universitätsbibliothek[199]
TB	Terabyte[200]
TEI	Text Encoding Initiative[201]
TEL	The European Library[202]
textMD	Techical Metadata for Text[203]
Tiff	Tagged Image File Format[204]
TRIM	Tower Records Information Management[205]
TSSP	Tessella plc[206]
U.S	United States
UDFR	Unified Digital Formats Registry[207]
UK	United Kingdon

UKOLN	UK Office for Library and Information Networking[208]
UPHEC	University of Portsmouth[209]
URL	Uniform Resource Locator[210]
URN	Uniform Resource Name[211]
URN:NBN	Uniform Resource Name: National Bibliographic Number[212]
USK	UnterhaltungssoftwareSelbstkontrolle[213]
UVC	Universal Virtual Computer[214]
VCS	Video Computer System[215]
VET	Valtion elokuvatarkastamo [216]
VIDEOMD	Video Metadata[217]
VRC	Videogame Rating Council[218]
WARC	Web ARCive File Format[219]
WP	Work Package[220]
XML	Extensible Markup Language [221]
XP	eXPerience[222]
XSLT	EXtensible Stylesheet Language Trasformation[223]

12 Bibliography

Anderson, D. P., (2007). Was the Manchester Baby conceived at Bletchley Park? British Computer Society: Electronic Workshops in Computing

Baker, T., Blanchi, C., Brickley, D., Duval, E., Heery, R., Johnston, P., et al. (2003). Principles of Metadata Registries (A White Paper of the DELOS Working Groon Registries)

Bearman, D. (1989). Archival methods. Pittsburgh, Pa.: Archives & Museum Informatics.

Bearman, D. (1999). Reality and Chimeras in the Preservation of Electronic Records. D-Lib Magazine, 5(4).

Blanchi, C., & Petrone, J. (2001). Distributed Interoperable Metadata Registry. D-Lib Magazine, 7(12).

Brown, A. (2008). White Paper: Representation Information Registries (Planets External Deliverable D7)

Buckland, M. K. (1997). What is a "Document"? Journal of the American Society for Information Science 48(9), 804-809.

Caplan, P. (2009). Understanding PREMIS. from http://www.loc.gov/standards/premis/understanding-premis.pdf

Caplan, P., & Guenther, R. (2005). Practical Preservation: The PREMIS Experience. Library Trends, 54(1).

CCSDS. (1999). Reference Model for an Open Archival Information System (OAIS) (Issue 1 Red Book). Washington, DC: CCSDS.

CCSDS. (2002). Reference Model for an Open Archival Information System (OAIS) (Issue 1 Blue Book). Washington, DC: CCSDS.

Chen, S.-S. (2001). The Paradox of Digital Preservation. EEE Computer(March 2001), 24-28.

CLOCKSS, (2008). CLOCKSS: A Trusted Community-Managed Archive. Retrieved 20th July 2009, from http://www.clockss.org/clockss/Home

Collins, R., & Murroni, C. (1996). New Media, New Policies. Cambridge: Polity Press.

Conway, P. (1990). Archival Preservation in a Nationwide Context. American Archivist, 53(2), 204-222.

Dappert, A., & Enders, M. (2008). Using METS, PREMIS and MODS for Archiving eJournals. D-Lib Magazine, 14(9/10).

Day, M. (1998). CEDARS: Digital Preservation and Metadata. Paper presented at the Sixth DELOS Workshop: Preservation of Digital Information. from http://www.ukoln.ac.uk/metadata/presentations/delos6/cedars.html

Day, M. (2001). Metadata for digital preservation: a review of recent developments. Paper presented at the ECDL2001, 5th European Conference on Research and Advanced Technology for Digital Libraries. from http://www.ukoln.ac.uk/metadata/presentations/ecdl2001-day/paper.html

DCMI, (2003). Dublin Core Metadata Element Set, Version 1.1: Reference Description. from http://dublincore.org/documents/2003/06/02/dces/

Deutsche Nationalbibliothek, (2009). Nestor - Network of expertise in long-term storage of digital resources. Retrieved 27th July 2009, from http://www.d-nb.de/eng/wir/projekte/nestor.htm

Diessen, R. J. v., & Werf, T. v. d. (2002). Authenticity in a Digital Environment. Den Haag: Koninkklijke Bibliotheek.

DiGA, e. V. (2003). Projekt Metadaten: Ein Blick auf verschiedene Softwarearchive und deren Datenstrukturen.

Digital Archive Metadata Elements. (2003). (OCLC Digital Archive System Guides). Dublin, Ohio, USA

Dobratz, S., & Neuroth, H. (2004). nestor: Network of Expertise in Long-term STOrage of Digital Resources - A Digital Preservation Initiative for German. D-Lib Magazine, 10(4).

Exon, M. (1995). Long-Term Management Issues in the Preservation of Electronic Information. Paper presented at the 2nd National Preservation Office Conference: Multimedia Preservation - Capturing the Rainbow. from http://www.nla.gov.au/niac/meetings/npo95me.html

Gartner, R. (2008). Metadata for digital libraries: state of the art and future directions. Bristol: Technology & Standards Watch

Gavrel, S. (1986). Preserving Machine-Readable Archival Records: A Reply to John Mallinson. Archivaria, 22 (Summer), 153-155.

Geser, G., Kasteren, J. v., Kop, P., Pereira, J., Ross, S., Steemson, M., et al. (2002). Integrity and Authenticity of Digital Cultural Heritage Objects: Thematic Issue 1. Austria.

Gesetz uber Die Deutsche Nationalbibliothek, (2006).

Goldman, E., (2004 January 7th). Warez Trading and Criminal Copyright Infringement. SSRN Retrieved July 22nd 2009, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=487163

Gooding, P., & Terras, M. (2008). 'Grand Theft Archive': A Quantitative Analysis of the State of Computer Game Preservation. International Journal of Digital Curation, 3(2), 19-41.

Heery, R., & Wagner, H. A. (2002). A Metadata Registry for the Semantic Web. D-Lib Magazine, 8(5), 15-17.

Holdsworth, D., & Wheatley, P., (2001). Emulation, preservation, and abstraction. from http://cedarsgw.leeds.ac.uk/CAMiLEON/dh/ep5.html

Hoorens, S., Rothenberg, J., van Oranje, C., van der Mandele, M., & Levitt, R. (2007). Addressing the uncertain future of preserving the past: Towards a robust strategy for digital archiving and preservation: Rand Europe, prepared for the Koninklijke Bibliotheek. (R. Europe)

Huth, K. (2004a). How to Preserve Computergames in Archives and Libraries? The Metadata Concept. Paper presented at the Wizards of OS 3. The Future of the Digital Commons. from http://www.wizards-of-os.org/archiv/wos_3/programm/panels/freier_content/offene_archive_ii_games/karsten_huth/pdf.html

Huth, K. (2004b). Probleme und Lösungsansätze zur archivierung von computerprogrammen - am beispeil der software des ATARI VCS 2600 und des C64. Humboldt-Universität Zu Berlin, Berlin.

Janée, G., Mathena, J., & Frew, J. (2008). A Data Model and Architecture for Long-term Preservation. Paper presented at the JCDL. from http://www.ngda.org/research/Tech%20Arch/jcdl-paper.pdf

Jones, M. (2002). The Cedars Project Library and Information Research News 26(84), 11-16.

KB, DNB, & SUB. (2008). DPFuse Report (Confidential. All rights are reserved by Deutsche National Bibliothek (DNB), the Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) and the National Library of the Netherlands (KB))

Kenney, A. R., Entlich, R., Hirtle, P. B., McGovern, N. Y., & Buckley, E. L. (2006). E-Journal Archiving Metes and Bounds: A Survey of the Landscape. British Library Conference Centre. (C. o. L. a. I. Resources)

Khong, D. W. K. (2007). Orphan Works, Abandonware and the Missing Market for Copyrighted Goods. International Journal of Law and Information Technology 15(1), 54-89.

Lavoie, B. F. (2004). Implementing Metadata in Digital Preservation Systems: The PREMIS Activity. D-Lib Magazine, 10(4).

Lavoie, B. F., & Gartner, R. (2005). Preservation Metadata: Online Computer Library Center Inc.(OCLC)

Library of Congress, (2007). METS Profiles. Retrieved 27th July 2009, from http://www.loc.gov/standards/mets/mets-profiles.html

Library of Congress, (2009). METS: An Overview & Tutorial. Retrieved 22nd July 2009, from http://www.loc.gov/standards/mets/METSOverview.v2.html

Lundy, M. W. (2008). Provenance evidence in bibliographic records: demonstrating the value of best practices in special collections cataloging. (Report): Library Resources & Technical Services.

Lupovici, C., & Masanès, J. (2000). Metadata for the Long-term Preservation of Electronic Publications. Den Haag: Koninklijke Bibliotheek

Lyman, P., & Varian, H. R., (2000). How Much Information? Retrieved 7th July 2009, from http://www.sims.berkeley.edu/how-much-info

Lyman, P., & Varian, H. R., (2003). How Much Information? Retrieved 7th July 2009, from http://www.sims.berkeley.edu/how-much-info-2003

Mallinson, J. C. (1986). Preserving Machine-Readable Archival Records for the Millenia. Archivaria, 22 (Summer), 147-152.

Marcum, D., & Friedlander, A. (2003). Keepers of the Crumbling Culture: What Digital Preservation Can Learn from Library History. D-Lib Magazine, 9(5).

McCargar, V. (2005). No Pain, No Metadata. The Seybold Report, 5(6).

McLeod, R., Wheatley, P., & Ayris, P. (2006). Lifecycle information for e-literature: full report from the LIFE project (Research report). London, UK

Metadata Standards Framework –Preservation Metadata (Revised). (2003). (Technical Paper): National Library of New Zealand

NISO. (2004). Understanding Metadata. Bethesda, MD: National Information Standards Organization

NLA, (1997 October 18th 2006). PANDORA Logical Data Model, Version 2. Retrieved 21st July 2009, 2009, from http://pandora.nla.gov.au/ldmv2.html#data

NLA, (1999). Preservation Metadata for Digital Collections. Retrieved 21st July 2009, 2009, from http://www.nla.gov.au/preserve/pmeta.html

NLA, (2001 February 20th 2009). PANDORA Digital Archiving System (PANDAS). Retrieved 21st July 2009, 2009, from http://pandora.nla.gov.au/pandas.html

NLNZ. (2003). Metadata Standards Framework –Preservation Metadata (Revised)

OCLC. (2006). OCLC Digital Archive Preservation Policy and Supporting Documentation (Policy Document). Dublin, Ohio: Online Computer Library Center

OCLC/RLG. (2001a). Preservation Metadata for Digital Objects: A Review of the State of the Art (A White Paper by the OCLC/RLG Working Group on Preservation Metadata)

OCLC/RLG. (2001b). Preservation Metadata for Digital Objects: A Review of the State of the Art.

OCLC/RLG. (2002). Preservation Metadata and the OAIS Information Model A Metadata Framework to Support the Preservation of Digital Objects. Dublin, Ohio: Online Computer Library Center, Inc. (OCLC)

OCLC/RLG. (2004). Implementing Preservation Repositories For Digital Materials: Current Practice And Emerging Trends In The Cultural Heritage Community (A Report by the PREMIS Working Group)

OCLC/RLG. (2005). The PREMIS Data Dictionary version 1.0.

Owen, J. M., Coutts, M., & Veling, S. (2001). Cedars Summative Evaluation Final Report

PREMIS, (2007). PREMIS Implementation Registry. Retrieved 23rd July 2009, from http://www.loc.gov/standards/premis/premis-registry.php

Preserving Digital Information. (1996). (Report of the Task Force on Archiving of Digital Information): The Commission on Preservation and Access and The Research Libraries Group

President and Fellows of Harvard College, (2006). Andrew W. Mellon Foundation Awards $600,000 to HUL for Global Digital Format Registry. from http://hul.harvard.edu/gdfr/mellon.html

Procter, M., & Cook, M. (2000). Manual of Archival Description (British Library R&D Report No. 5965 ISBN 978-0566082580): Gower Publishing Ltd

Reich, V., & Rosenthal, D. S. H. (2001). LOCKSS: A Permanent Web Publishing and Access System. D-Lib Magazine, 7(6).

Rothenberg, J. (1999). Avoiding technological quicksand : finding a viable technical foundation for digital preservation (No. 1887334637 (acid-free paper)). Washington, DC: Council on Library and Information Resources

Rothenberg, J. (2000). An experiment in using emulation to preserve digital publications. Den Haag: Koninklijke Bibliotheek

Rothenberg, J., & Bikson, T. K. (1999). Carrying Authentic, Understandable and usable Digital Records Through Time (Final Report): RAND Corporation

SHAMAN. Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN) Project Retrieved 23rd July 2009, from http://shaman-ip.eu/shaman/node/44

Smith, A. (2000a). Library Collections Online. In Collections, Content and the Web. (pp. 34-39). Washington, DC: Council on Library and Information Resources.

Smith, A. (Ed.). (2000b). Authentication of Digital Objects: Lessons from a Historian's Research. Washington D.C.: Council on Library and Information Resources.

Smith, M. (2005). Eternal bits: How can we preserve digital files and save our collective memory? IEEE Spectrum, 42(7).

Steinke, T. (2005). LMER: Long-term preservation Metadata for Electronic Resources (Reference description: Translation into English: Dr. Thomas Wollschläger): Die Deutsche Bibliothek

Stone , A., & Day, M. (1999). Cedars Preservation Metadata Elements: UKOLN: The UK Office for Library and Information Networking, University of Bath

Unified Digital Formats Registry (UDFR): Proposal and roadmap. (2009).

van der Hoeven, J., Lohman, B., & Verdegem, R. (2007). Emulation for Digital Preservation in Practice: The Results. International Journal of Digital Curation, 2(2).

van der Werf, T. (2000). The Deposit System for Electronic Publications: A Process Model. Den Haag.

van der Werf-Davelaar, T. (1999). Long-term Preservation of Electronic Publications: The NEDLIB project. D-Lib Magazine, 5(9).

Verheul, I. (2006). Networking for Digital Preservation: Current Practice in 15 National Libraries. . München: Koninklijke Bibliotheek / IFLA / Saur.

Waters, D., & Garrett, J. (1996). Preserving Digital Information: Report of the Task Force on Archiving of Digital Information [Electronic Version], 71. Retrieved May 1 1996, from http://www.clir.org/pubs/reports/pub63watersgarrett.pdf

Wheatley, P. (2001). Migration - a CAMiLEON discussion paper. Ariadne(29).

Winget, M. (2005). Like a Wave Upon the Sand: Digital Preservation of Variable Media Art. Paper presented at the (im)permanence: Cultures in/out of time conference. from http://www.ischool.utexas.edu/~megan/research/Winget_DigPres.pdf

Woodyard, D., & Bradley, K. (2000). Preservation Metadata for Digital Collections. Paper presented at the Metadata for long term preservation in NEDLIB. from http://www.nla.gov.au/openpublish/index.php/nlasp/article/view/1344/1628

Woodyard-Robinson, D. (2006). Common Use Cases for Preservation Metadata

Woodyard-Robinson, D. (2007). Implementing the PREMIS data dictionary: a survey of approaches

13 Endnotes

1 http://www.lockss.org/lockss/Home

2 A programming language for quantum computers. See http://tph.tuwien.ac.at/~oemer/qcl.html

3 A very similar idea underpins the KEEP project. See section 2.6 of this report.

4 Metadata for Images in XML

5 Text Encoding Initiative

6 Audio Technical Metadata Extension Schema

7 Video Technical Metadata Extension Schema

8 Open Digital Rights Language

9 This followed a number of previous 'draft' documents, the earliest of which was issued in May 1999. See CCSDS (1999).

10 Many of the preservation metadata schemas discussed in this report are said to be OAIS-compliant, including CEDARS, NEDLIB, LMER, PREMIS and METS.

11 The PREMIS Working group polled 49 institutions (28 Libraries, 7 Archives and 14 'Other') during November 2003. It should be kept in mind that surveys are indicative 'snapshots' at best and, particularly when viewed in isolation, cannot be taken as definitive.

12 CURL is a consortium of UK research libraries whose mission is "to promote, maintain and improve library resources for research in universities."

13 Accessed 21st July 2009

14 The government of Scotland

15 http://www.natlib.govt.nz/catalogues/library-documents/preservation-metadata-revised

16 METS http://www.loc.gov/standards/mets/

17 http://www.loc.gov/standards/mets/METSOverview.v2.html

18 Examples of METS profiles are available at: http://www.loc.gov/standards/mets/profile_docs/

19 The METS implementation registry http://www.loc.gov/standards/mets/mets-registry.html#bnbnd

20 http://www.liwa-project.eu/

21 http://www.parse-insight.eu/

22 http://www.protage.eu/

23 http://www.prestoprime.eu/

24 http://shaman-ip.eu/shaman/

25 http://www.casparpreserves.eu/

26 http://www.digitalpreservationeurope.eu/

27 http://www.p2p-fusion.org/

28 http://www.planets-project.eu/

29 http://prestospace.org/

30 http://www.europeana.eu/portal/

31 http://www.d-nb.de/eng/wir/projekte/nestor.htm

32 http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

33 http://www.gdfr.info/index.html

34 Digital library program manager in Harvard University Library's Office for Information Systems

35 http://www.udfr.org/

36 http://www-05.ibm.com/nl/dias/preservatiomanager.html

37 http://www.tessella.com/

38 http://www.bnf.fr/pages/zNavigat/frame/version_anglaise.htm?ancre=english.htm

39 http://www.d-nb.de/eng/index.htm

40 http://www.kb.nl/index-en.html

41 http://www.computerspielemuseum.de/

42 http://www.kb.nl/bst/beleid/missie-en.html

43 http://www.d-nb.de/eng/wir/ueber_dnb/dnb_im_ueberblick.htm

44 http://www.bnf.fr/pages/zNavigat/frame/version_anglaise.htm?ancre=english.htm

45 http://www.europeana.eu/portal/

46 The CSM are also part of the nestor competence network.

47 CSM are external contributors to Planets.

48 http://www.nla.gov.au/padi/topics/67.html#Ger

49 Created by J-Ph Humblot

50 Created by V. Joguin see http://www.joguin.com/

51 http://dioscuri.sourceforge.net/

52 http://kopal.langzeitarchivierung.de/

53 http://public.ccsds.org/publications/archive/650x0b1.pdf

54 http://www-05.ibm.com/nl/dias/

55 http://kopal.langzeitarchivierung.de/index_koLibRI.php.en

56 http://hul.harvard.edu/jhove/ & http://www.jstor.org/

57 http://kopal.langzeitarchivierung.de/index_objektspezifikation.php.en

58 http://www.persistent-identifier.de/?lang=en

59 http://www.daffodil.de

60 http://www.d-nb.de/netzpub/ablief/metadiss.htm#metadiss

61 http://www.d-nb.de/eng/standards/pdf/ref_xmetadiss_v1-3.pdf

62 http://www.mcu.es/bibliotecas/CE/jppd/ficheros/reinhard_altenhoner_jpd.pdf

63 www.loc.gov/standards/mets

64 http://opus.kobv.de/zib/volltexte/2005/893/pdf/ZR-05-60.pdf

65 http://www.d-nb.de/standardisierung/formate/mabxml.htm See also http://www.ib.hu-berlin.de/~voj/MABxmlTools/

66 http://www.loc.gov/marc/

67 http://www.loc.gov/standards/marcxml/

68 http://www.editeur.org/onix.html

69 http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

70 http://www.nla.gov.au/openpublish/index.php/nlasp/article/viewArticle/1314/1600

71 http://www.zavatar.de/

72 https://www.media-control.com/

73 http://www.retro-games.co.uk/atari/atari_vcs.htm

74 http://www.commodore.ca/products/c64/commodore_64.htm

75 http://www.mobygames.com/home

76 Total verified at http://www.mobygames.com/browse/games/full,100/ [accessed 19th July 2009]

77 http://uk.gamespot.com/

78 The genre field shares similarities with MobyGames although there are instances of alternate categorisation of media and the full list of categories are not available.

79 This is reduced for console games which simply carry the basic platform name.

80 http://www.abandonia.com/

81 http://www.homeoftheunderdogs.net/

82 An externally-linking field directing the user to other community preservation sites to obtain actual media objects

83 http://www.adese.es/web/asociacion_adese.asp

84 http://ahds.ac.uk/

85 http://public.ccsds.org/publications/archive/650x0b1.pdf

86 http://www.bnf.fr/

87 http://www.c64.com/

88 http://www.casparpreserves.eu/

89 http://www.cce.org.pt/ENGLISH1.htm

90 http://public.ccsds.org/default.aspx

91 http://www.webopedia.com/TERM/C/CD_ROM.html

92 http://www.webopedia.com/TERM/C/CD_RW_disk.html

93 http://www.webarchive.org.uk/ukwa/target/99695/

94 http://en.wikipedia.org/wiki/Cpu

95 http://www.computerspielemuseum.de/

96 http://www.nla.gov.au/padi/metafiles/resources/42.html

97 http://fclaweb.fcla.edu/content/digital-preservation

98 http://www.dcc.ac.uk/about/

99 http://dublincore.org/

100 http://www-05.ibm.com/nl/dias/

101 http://www.digitalgamearchive.org/home.php

102 http://dgrc.ncsu.edu/

103 http://www.din.de/cmd?level=tpl-home&languageid=en

104 http://public.ccsds.org/publications/archive/650x0b1.pdf

105 http://en.wikipedia.org/wiki/DJCTQ

106 http://www.diglib.org/

107 http://www.d-nb.de/

108 http://en.wikipedia.org/wiki/DOS

109 http://www.dpconline.org/graphics/index.html

110 http://www.driver-repository.eu/

111 http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html

112 http://www.w3schools.com/DTD/default.asp

113 http://en.wikipedia.org/wiki/Dvd

114 http://www.webopedia.com/TERM/D/DVD_RW.html

115 http://en.wikipedia.org/wiki/E-book

116 http://www.economy-point.org/e/european-article-number.html

117 http://ec.europa.eu/information_society/activities/econtentplus/projects/cult/edlnet/index_en.htm

118 http://www.egdf.eu/

119 http://www.ukoln.ac.uk/services/elib/

120 http://www.esrb.org/index-js.jsp

121 http://fclaweb.fcla.edu/

122 http://www.fedora-commons.org/about

123 http://www.engin.umd.umich.edu/CIS/course.des/cis400/fortran/fortran.html

124 http://www.fpb.gov.za/

125 http://en.wikipedia.org/wiki/File_Transfer_Protocol

126 http://www.gdfr.info/about.html

127 http://www.homeoftheunderdogs.net/about.php

128 http://www.ibm.com/us/en/

129 http://www.impact-project.eu/

130 http://www.iso.org/iso/home.htm

131 http://www.issn.org/2-22636-All-about-ISSN.php

132 http://hul.harvard.edu/jhove/

133 http://www.jisc.ac.uk/

134 http://www.joguin.com/

135 http://www.jstor.org/page/info/about/organization/index.jsp

136 http://www.kb.nl/index-en.html

137 http://www.keep-project.eu/ezpub2/index.php

138 http://www.kmrb.or.kr/kmrb_2008/english/

139 http://kopal.langzeitarchivierung.de/index_koLibRI.php.en

140 http://kopal.langzeitarchivierung.de/index.php.de

141 http://www.liwa-project.eu/

142 http://www.d-nb.de/eng/standards/lmer/lmer.htm

143 http://www.lockss.org/lockss/Home

144 http://www.d-nb.de/standardisierung/formate/mab.htm

145 http://journals.sfu.ca/archivar/index.php/archivaria/article/viewFile/11770/12]

146 http://mamedev.org/

147 http://www.loc.gov/marc/

148 http://www.mess.org/

149 http://www.loc.gov/standards/mets/

150 http://www.loc.gov/standards/mix/

151 http://www.mpeg.org/

152 http://www.microsoft.com/en/us/default.aspx

153 http://www.nas.gov.uk/

154 http://pandora.nla.gov.au/pan/21336/20031011-0000/www.nla.gov.au/abn/nbdreports.html

155 http://www.neh.gov/projects/ndnp.html

156 http://www.nla.gov.au/padi/metafiles/resources/130.html

157 http://www.langzeitarchivierung.de/index.php

158 http://www.niso.org/home

159 http://www.nla.gov.au/

160 http://www.natlib.govt.nz/

161 http://www.bl.uk/npo/aboutkey.html

162 http://www.openarchives.org/pmh/

163 http://public.ccsds.org/publications/RefModel.aspx

164 http://www.oclc.org/uk/en/default.htm
165 http://en.wikipedia.org/wiki/Optical_character_recognition
166 http://odrl.net/
167 http://www.oflc.gov.au/
168 http://www.editeur.org/onix.html
169 http://en.wikipedia.org/wiki/Online_public_access_catalog
170 http://en.wikipedia.org/wiki/Operating_system
171 http://en.wikipedia.org/wiki/Peer-to-peer
172 http://www.nla.gov.au/nla/staffpaper/2004/koerbin2.html
173 http://pandora.nla.gov.au/about.html
174 http://www.parse-insight.eu/
175 http://www.britannica.com/EBchecked/topic/452928/personal-computer
176 http://en.wikipedia.org/wiki/Pdf
177 http://public.ccsds.org/publications/archive/650x0b1.pdf
178 http://www.pegi.info/en/index/
179 http://www.persistent-identifier.de/english/204-examples.php
180 http://www.planets-project.eu/
181 http://www.site.uottawa.ca:4321/oose/index.html#person-month
182 http://en.wikipedia.org/wiki/Powerpoint
183 http://www.oclc.org/research/projects/pmwg/default.htm
184 http://prestospace.org/
185 http://www.protage.eu/
186 http://purl.org/
187 http://tph.tuwien.ac.at/~oemer/qcl.html
188 http://en.wikipedia.org/wiki/Random-access_memory
189 http://www.w3.org/RDF/
190 http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformati
191 http://www.oclc.org/programs/about/default.htm
192 http://www.mobygames.com/genre/sheet/role-playing-rpg/
193 http://www.rsac.org/
194 http://en.wikipedia.org/wiki/Science_fiction
195 http://shaman-ip.eu/shaman/
196 http://public.ccsds.org/publications/archive/650x0b1.pdf
197 http://en.wikipedia.org/wiki/Small_and_medium_enterprises
198 http://bibnum.bnf.fr/spar/index.html

199 http://www.sub.uni-goettingen.de/

200 http://www.webopedia.com/TERM/t/terabyte.html

201 http://www.tei-c.org/index.xml

202 http://search.theeuropeanlibrary.org/portal/en/index.html

203 http://www.loc.gov/standards/textMD/

204 http://www.ietf.org/rfc/rfc2302.txt

205 http://pandora.nla.gov.au/pan/77653/20071017-1101/www.nt.gov.au/dcis/nta/recordkeeping/advice3.html

206 http://www.tessella.com/

207 http://www.gdfr.info/udfr.html

208 http://www.ukoln.ac.uk/

209 http://www.port.ac.uk/

210 http://en.wikipedia.org/wiki/Uniform_Resource_Locator

211 http://www.ietf.org/rfc/rfc3406.txt

212 http://www.kb.se/english/about/projects/digital/urn-nbn/

213 http://www.usk.de/

214 http://www.nla.gov.au/padi/topics/492.html

215 http://en.wikipedia.org/wiki/Atari_2600

216 http://www.vet.fi/

217 http://www.loc.gov/rr/mopic/avprot/videoMD_v8.xsd

218 http://en.wikipedia.org/wiki/Videogame_Rating_Council

219 http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

220 http://en.wikipedia.org/wiki/Work_package

221 http://www.w3.org/XML/

222 http://www.microsoft.com/windows/windows-xp/default.aspx

223 http://www.w3schools.com/xsl/