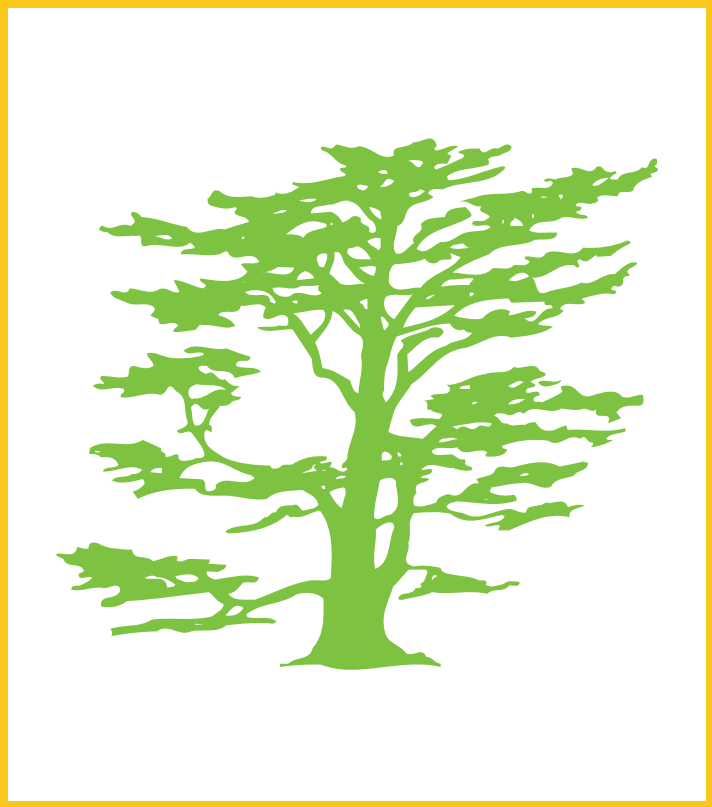
Cedars Guide to



Preservation Metadata

Date: March 2002

url: http://www.leeds.ac.uk/cedars/guideto/metadata/

Cedars Guide to Preservation Metadata

CONTENTS

Foreword

The Cedars (**C**URL **E**xemplars in **D**igital **AR**chive**s**) Project ran from April 1998 until March 2002. Funded by JISC (the Joint Information Systems Committee of the UK higher education funding councils), as part of its Electronic Libraries (eLib) Programme, Cedars was the only project in the programme to focus on digital preservation. The project was a collaboration between three CURL institutions, the universities of Leeds, Oxford, and Cambridge. As this is such a new and rapidly developing area, and of crucial importance to the future of scholarly research, it was felt important from the outset to ensure that there were mechanisms in place to share the work of Cedars with a wider audience. The series of Cedars Guides is designed to disseminate achievements of the project in five major areas: Preservation Metadata; Intellectual Property Rights; Collection Management; Technical Strategies; and the Digital Archiving Prototype. The guides are available in printed form and are also available from the project website at http://www.leeds.ac.uk/cedars/. During the course of the Cedars project, a great deal of work on digital preservation has been undertaken around the world and much progress has been made in understanding the complex issues involved. Cedars has maintained close contact with many of these activities and has forged a close working relationship with them. This wider perspective is reflected in the Cedars Guides. This series of guides is aimed principally at librarians, who need to plan for and manage, increasing quantities of digital resources. However, we believe that they will provide a useful source of reference for anyone interested in digital materials, including creators of digital content, records managers, and archivists. No detailed technical knowledge is assumed though a broad awareness of the issues would be helpful to the understanding of the text.

Clare Jenkins

Cedars Project Director

Executive Summary

This document provides some background on preservation metadata for those interested in digital preservation. It first attempts to explain why preservation metadata is seen as an essential part of most digital preservation strategies. It then gives a broad overview of the functional and information models defined in the *Reference Model for an Open Archival Information System (OAIS)* and describes the main elements of the Cedars outline preservation metadata specification. The next sections take a brief look at related metadata initiatives, make some recommendations for future work and comment on cost issues. At the end there are some brief recommendations for collecting institutions and the creators of digital content followed by some suggestions for further reading.

1 Audience and Purpose

This document is intended to provide a brief introduction to current preservation metadata developments and introduce the outline metadata specification produced by the Cedars project. It is aimed in particular at those who may have responsibility for digital preservation in the UK further and higher education community, e.g. senior staff in research libraries and computing services. It should also be useful for those undertaking digital content creation (digitisation) initiatives, although it should be noted that specific guidance on this is available elsewhere (e.g., Kenney & Rieger, 2000; Grout & Ingram, 2001; UKOLN, 2001). The guide may also be of interest to other kinds of organisations that have an interest in the long-term management of digital resources, e.g. publishers, archivists and records managers, broadcasters, etc.

The document aims to provide:

- A rationale for the creation and maintenance of preservation metadata to support digital preservation strategies, e.g. migration or emulation.
- An introduction to the concepts and terminology used in the influential ISO Reference Model for an Open Archival Information System (OAIS).
- Brief information on the Cedars outline preservation metadata specification and the outcomes of some related metadata initiatives.
- Some notes on the cost implications of preservation metadata and how these might be reduced.

2 Background

Digital preservation has been defined as "the planning, resource allocation, and application of preservation methods and technologies necessary to ensure that digital information of continuing value remains accessible and usable" (Hedstrom, 1999, p. 189). The reasons why preserving digital information is difficult are technological, related to things like relatively short media lifetimes, obsolete hardware and software, and defunct Web sites (Chen, 2001, p. 24). Proposed solutions are partly technological, partly organisational. Various preservation strategies have been proposed; for example, there has been much recent discussion about the relative benefits of migration and emulation (e.g., Russell, 2000, pp. 143-147). Alternative approaches might include keeping museums of obsolete hardware or the relatively expensive data recovery programmes that are sometimes known as 'digital archaeology' (e.g. Ross & Gow, 1999). Regardless of which particular strategy is adopted, long-term preservation will depend upon the generation and maintenance of data that describe the digital information being preserved and to enable its interpretation. This data can be viewed as metadata, usefully defined as "structured information that describes and/or allows us to find, manage, control, understand or preserve other information over time" (Cunningham, 2000, p. 9). It is often envisaged that metadata will be part of the wrapping (or encapsulation) of digital objects and that such objects will effectively be self-documenting (Waugh, *et al.*, 2000, p. 175).

At heart, preservation metadata is all of the various types of data that will allow the re-creation and interpretation of the structure and content of digital data that has been preserved (Ludäsher, Marciano & Moore, 2001). Defined in this way, it is clear that preservation metadata needs to support a number of related, but distinct, functions. Lynch (1999), for example, says that within a digital repository, "metadata accompanies and makes reference to each digital object and provides associated descriptive, structural, administrative, rights management, and other kinds of information." The wide range of functions that preservation metadata is aimed to fulfil means that defining metadata standards is not a simple task and that most of the currently published schemas are relatively complex. The situation is complicated further by the perception that different kinds of metadata will be required to support different digital preservation strategies or digital information types.

At the time the Cedars project proposal was being put together, there was an awareness of the perceived importance of preservation metadata, but there was no existing 'standard' that could be adopted by the project for use in its demonstrator services. The project bid proposed, therefore, that Cedars would produce a metadata specification. Before work on developing the specification started, UKOLN undertook a review of preservation metadata initiatives for Cedars (Day, 1998). This described some of the more prominent initiatives in the areas of recordkeeping metadata, digital imaging and other areas. These included Bearman & Sochats (1996) influential *Metadata requirements for evidence* and the logical data model developed for the National Library of Australia's PANDORA (Preserving and Accessing Networked DOcumentary Resources of Australia) project. The report also briefly appraised the then latest draft of the *Reference Model for an Open Archival Information System* (OAIS), and confirmed that its 'taxonomy of archival information object classes' was of interest to Cedars in developing its metadata specification.

3 The Cedars Metadata Specification

The Cedars specification had two main aims. Firstly to develop a scheme that could be used within the Cedars demonstrator services, secondly as a contribution to international efforts at standardisation on preservation metadata. Work on developing the Cedars metadata specification started in early 1999. An initial draft (for expert comment) was published in January 2000, and was broadly organised according to the information model provided in the influential *Reference Model for an Open Archival Information System (OAIS)* published by the Consultative Committee on Space Data Systems (CCSDS). Before attempting to describe the Cedars schema, we will need to describe some of the features of the OAIS document.

3.1 *The Reference Model for an Open Archival Information System*

The OAIS model aims to provide a common framework that can be used to help understand archival challenges and especially those that relate to digital information. This is the model's real value: providing a high-level common language that can facilitate discussion across the different communities interested in digital preservation. The document defines a high-level reference model for an OAIS, which is defined as an organisation of people and systems that have "accepted the responsibility to preserve information and make it available for a Designated Community" (CCSDS, 2001, p. 1-11).

The OAIS model has a much wider scope than metadata. It defines both a functional model and an information model. The functional model outlines the range of functions that would need to be undertaken by a repository, and defines in more detail those functions described within the OAIS specification as access, administration, archival storage, data management, ingest and preservation planning (Fig. 1). The information model defines the broad types of information (or metadata) that would be required in order to preserve and access the information stored in a repository. However, it is important to realise that the OAIS standard is a reference model, not a detailed specification for any implementation based on it. All of the different communities interested in digital preservation will have to apply the model (including the information model) in their own particular contexts, both organisational and technical.

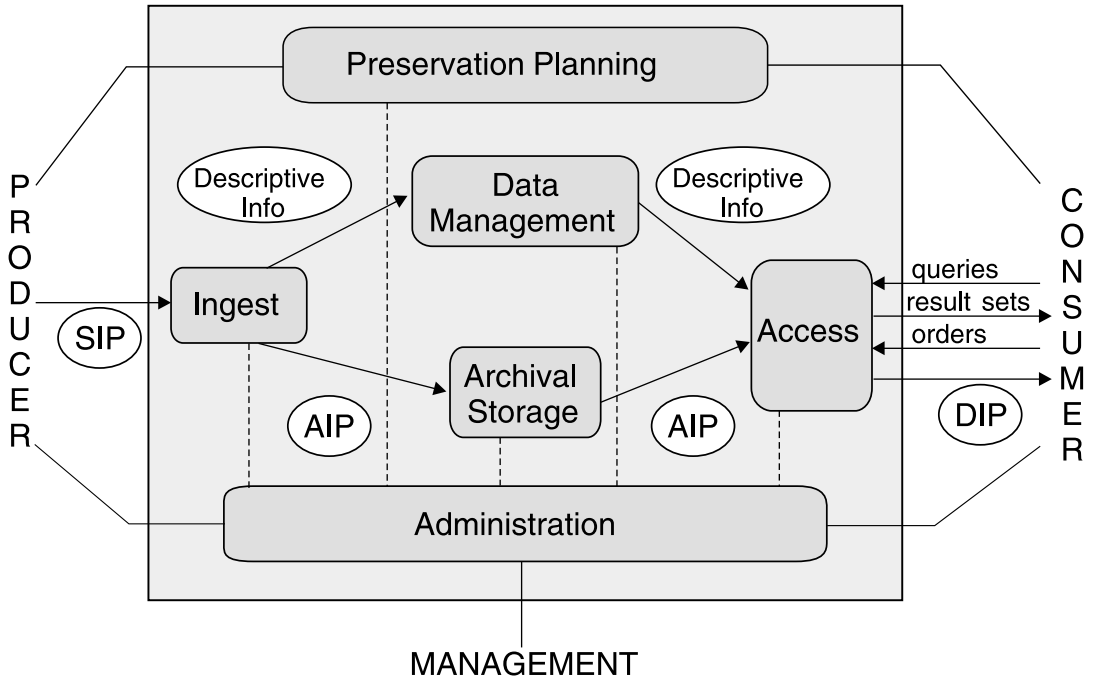


Figure 1: OAIS Functional Entities. Source: CCSDS (2001), Fig. 4-1

It is important to remember that the OAIS is a reference model and not a blueprint for an archive implementation.

3.2 *The OAIS information model*

The OAIS information model defines a number of different Information Objects that cover the various types of information required for long term preservation. A basic assumption of the model is that all Information Objects are composed of a Data Object -typically a sequence of bits for digital data - and the Representation Information that would permit the full interpretation of the Data Object into meaningful information (CCSDS, 2001, p. 4-19). The OAIS model defines four distinct Information Objects.

- *Content Information* - the information that requires preservation.
- *Preservation Description Information (PDI)* - any information that will allow the understanding of the Content Information over an indefinite period of time.
- *Packaging Information* - the information that binds all other components into a specific medium.
- *Descriptive Information* - information that helps users to locate and access information of potential interest. This could be based on information that is stored as part of the PDI, but is logically distinct.

The OAIS information model sub-divides the PDI into four distinct groupings, based on categories discussed in the report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access (CPA) and the RLG (CCSDS, 2001, p. 4-28). The task force wrote that "in the digital environment, the features that determine information integrity and deserve special attention for archival purposes include the following: content, fixity, reference, provenance and context" (Garrett & Waters, 1996). Accordingly, the OAIS taxonomy divides PDI into four: Reference Information, Context Information, Provenance Information and Fixity Information.

- *Reference Information* - any information that helps to identify and describe the Content Information. This would specifically include the unique identifiers used to identify the Content Information within the repository and, where appropriate, basic descriptive-type information that could be extracted to form part or all of the Descriptive Information.
- *Context Information* - defined as information that "documents the relationships of the Content Environment to its environment ... why the Content Information was created, and how it relates to other Content Information objects existing elsewhere" (CCSDS, 2001, p. 4-28). The CPA/RLG report suggested that 'context' should include information on the technical context of a digital object (Garrett & Waters, 1996), but some of this information is assigned in the OAIS model to the Packaging Information. (CCSDS, 2001, p. B-1).
- *Provenance Information* - information that documents the history of the Content Information. This might include information on its source or origin, any changes that may have taken place (e.g. migrations), and a record of the chain of custody. The CPA/RLG report says that the "assumption underlying the principle of provenance is that the integrity of an information object is partly embodied in tracing from where it came" (Garrett & Waters, 1996).
- *Fixity Information* - refers to any information that documents the particular authentication mechanisms in use within a particular repository. The CPA/RLG report comments that if the content of an object is "subject to change or withdrawal without notice, then its integrity may be compromised and its value as a cultural record would be severely diminished" (Garrett & Waters, 1996). Changes can either be deliberate or unintentional, but either type would adversely effect the integrity of Content Information.

The OAIS model also defines a conceptual structure for Information Packages. This is viewed as a container that logically encapsulates Content Information and its associated PDI within a single Data Object. Information Packages are defined for submission (SIP), archival storage (AIP) and dissemination (DIP). Of these, the Archival Information Package (AIP) is the most important for digital preservation, as it contains "all of the qualities needed for permanent, or indefinite, Long Term Preservation or a designated Information Object" (CCSDS, 2001, p. 4-33).

3.3 *The Cedars metadata specification*

The Cedars project team took the OAIS information model and used it as a broad framework for an outline preservation metadata specification (Russell, *et al.*, 2000). It is an outline specification because in many cases it only defines the highest levels of the metadata scheme that would be required for any implementation. Also, elements and sub-elements are not specified as being 'mandatory' or 'optional,' but just given a significance level. In accordance with the OAIS's Information Package model, the project team envisaged that resources (Content Information) would be packaged together with its metadata (PDI). The specification focused on defining both the Representation Information that would enable the Content Information Data Object to be understood (e.g. Holdsworth & Sergeant, 2000) and the Content Information's associated PDI. Less consideration was given to the specific Representation Information that would be required for the PDI Data Object, or to Packaging or Descriptive Information.

The Cedars project team was aware that the proposed metadata element set would not necessarily support all of the roles identified in the OAIS functional model, e.g. the administration or data management

functions. Despite this, however, it was recognised that some of the information provided as part of the Provenance Information could help support administrative functions like rights management. In fact, the Provenance Information defined in the Cedars outline specification contains a number of elements specific to rights management that goes well beyond the OAIS model's assumption that provenance is primarily concerned with supporting the integrity of a given Data Object. This reflects the difficulty of defining simple metadata schemes where the same information can be used by functionally different parts of a system.

A quick look at the hierarchical structure of the Cedars specification (Appendix A) demonstrates its basic dependence upon the OAIS information model. The first three levels of the hierarchy inherit the exact terminology and some of the definitions used in the OAIS model.

- The *Reference Information* section of the PDI has elements for a 'Resource Description' and a placeholder for any 'Existing Metadata.' The Cedars specification doesn't make any specific recommendations as to which elements would be included in the 'Resource Description,' but notes that any project-specific implementation would use an instantiation of the Dublin Core Metadata Element Set (DCMES). In a similar way, the precise way in which 'Existing Metadata' would be stored or utilised is not defined. In an operational repository, it is possible that at least some of the Descriptive Information and Reference Information would be generated automatically or extracted from metadata that already exists, e.g. in publishers databases or library catalogues.
- *Context Information* has one sub-element, referring to 'Related Information Objects.' This is supposed to specify information objects that are judged to have a significant relationship to the object being preserved. Again, what precise information would be required (e.g. an identifier, descriptive information, etc.) is not defined
- *Provenance Information* makes up the largest part of the Cedars metadata specification. The 'History of Origin' sub-section is intended to record the reasons why the object being preserved was created, its custody history before ingest, and to document why it is being preserved. This section also records technical information about the original technical environment of the object and any prerequisites with regard to software, operating systems, etc. A separate section on 'Management History' is supposed to keep information about the ingest process, and the policies and actions applied to objects since they were added to the repository. A final section on 'Rights Management' comprises a detailed set of sub-elements to help record and manage the intellectual property rights held in objects.
- *Fixity Information* contains a single sub-element, 'Authentication Indicator,' which is intended to record mechanisms used to ensure the digital object's authenticity, e.g. digital certificates or a checksum.

The Cedars outline metadata specification was developed firstly to help support the development of the project's demonstrator services and secondly as a contribution to international standardisation on preservation metadata. The specification tried not to make too many assumptions about the actual form of the digital objects being preserved or about the 'granularity' of specific objects. It was hoped that the specification would be applicable at any level of granularity, but the authors recognised that the specifics of implementation would be the responsibility of repositories. Also, the specification made no assumptions about which particular preservation strategy would be used, although this may have an impact on which particular elements would be required.

After publication of the outline specification, meetings were held in Birmingham and Cambridge to 'walk-through' the metadata element set with regard to specific resources. These raised many issues related to how the specification should be implemented and with regard to the organisation of metadata handling within a repository. This included questions about who would be responsible for generating this metadata and the relevant workflow.

4 Other Metadata Initiatives

At around about the same time as the Cedars project was developing its initial draft metadata specification, several other groups were beginning to develop and publish similar schemas. These originated in three main areas: deposit libraries thinking about their responsibilities with regard to digital information, libraries involved in digital content creation (digitisation) programmes and archivists developing systems for electronic recordkeeping.

4.1 *Deposit libraries*

Legal deposit libraries have an obvious interest in digital preservation. Many countries have already extended legal deposit legislation to cover digital publications - at least offline ones - and most of those that have not done so are actively considering taking this step (Muir, 2001). For this reason, some national libraries and other deposit libraries have recently begun to get involved in research and development activity related to digital preservation. Some of the more interesting developments in this area have related to Web preservation, examples being are the harvesting-based Swedish Royal Library's Kulturarw3 project (Arvidson, Persson & Mannerheim, 2000) and the Library of Congress's Minerva prototype (Arms, *et al*., 2001). Apart from Cedars, the main deposit library-based activity that has been involved in the development of preservation metadata element sets has been undertaken by the National Library of Australia (NLA) and a consortium of European libraries (and other organisations) in the NEDLIB project.

The National Library of Australia

The NLA has had a keen interest in digital preservation issues for a long time, demonstrated, for example, by its support and hosting of the PADI (Preserving Access to Digital Information) service. It has also been involved in developing frameworks for preservation that are, almost uniquely, based on practical experience (Webb, 2000). In 1996, the NLA established its PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) project, initially to provide an operational 'proof-of-concept' service. The PANDORA archive has since developed into a small but growing repository of selected Australian Web publications (http://pandora.nla.gov.au/). With regard to metadata, descriptive metadata for each object in the PANDORA archive is stored in the NLA's own library management system; individual items being identified by means of Persistent Uniform Resource Identifiers (PURLs). The project also developed a logical data model (based on entity-relationship modelling) to help identify the particular entities (or metadata) that would need to be supported by the repository.

The NLA also developed a specification of *Preservation Metadata for Digital Collections* (Phillips, *et al*., 1999). This was expressly based on an 'data output model,' i.e. it defined the information that a digital storage system would need to generate in order to facilitate the preservation management of digital content. The NLA metadata element set defined 25 high level elements (some with sub-elements) at three distinct levels of granularity: collection, object and sub-object (file). Unlike the Cedars specification, the NLA element set is not structured according to the OAIS information model, although it claims to have been influenced by it and other models. A mapping of the NLA elements to the OAIS information model has been published in the white paper published by the OCLC/RLG Working Group on Preservation Metadata (2001).

The NEDLIB project

The NEDLIB (Networked European Deposit Library) project ran from 1998 to 2000 and was funded by the European Commission as part of its Telematics Applications Programme. The project was based on a consortium of national libraries, publishers, information technology organisations and a national archive, all led by the National Library of the Netherlands. The project developed an architectural framework for what it called a deposit system for electronic publications (DSEP) that was broadly based on the OAIS model (Van der Werf, 2000).

As part of NEDLIB, the Bibliothèque nationale de France attempted to define the minimum metadata elements that would be necessary for preservation management (Lupovici & Masanès, 2000). Like the Cedars outline specification, the NEDLIB element set explicitly adopted the terminology and structure of the information model defined as part of the OAIS model. The element set, however, was much smaller than that proposed by Cedars (18 elements, 38 sub-elements) because it was focussed on only identifying 'core' (or mandatory) metadata elements. Unlike Cedars, NEDLIB was also primarily concerned with defining metadata that would address the problem of technological obsolescence, not with metadata for descriptive, administrative or legal purposes. The top-level metadata elements for preservation and description included ones for reference information (including identifiers assigned by the repository), fixity (e.g., a checksum) and change history. Information on specific hardware and software requirements, formats and applications are viewed as being part of the Representation Information.

4.2 *Digitisation initiatives*

The proliferation of cultural heritage-based digitisation programmes has meant that many institutions and projects now need to face the digital preservation problem at the planning stage. In fact, a consideration of digitised resources' future digital preservation needs are often now a condition of getting funding. For example, the standards document developed for the JISC's Distributed National Electronic Resource (DNER) says that projects must make "arrangements for long-term preservation and access with an appropriate repository," who will then "be responsible for implementing long-term preservation strategies and procedures" (Grout & Ingram, 2001). Similarly, a consideration of digital preservation issues forms an important part of the *Technical Standards and Guidelines* published for the UK New Opportunities Fund (NOF) digitisation of learning materials grant programme (UKOLN, 2001).

The existence of specialised digitisation centres and their long experience of creating digital resources (e.g., Smith, 2001) means that several metadata standards specific to digitisation initiatives have already been developed. For example, back in 1997, the RLG constituted a working group on the Preservation Issues of Metadata to help identify the kinds of information that would be required to manage a digital image master file over time. The final report of the working group defined sixteen metadata elements (RLG, 1998). A more complex metadata scheme was developed by the Making of America II (MOA2) testbed project (Hurley, *et al*., 1999), the general framework of which has recently been taken up in the Digital Library Federation's METS initiative.

The METS initiative

The Metadata Encoding & Transmission Standard (http://www.loc.gov/standards/mets/) initiative is attempting to provide an XML-based document format for encoding metadata to aid the management and exchange of digital library objects. The initiative has adapted the XML Document Type Definition developed by MOA2 to create an XML schema. The schema defined by the METS initiative separates metadata into four sections. These are 'descriptive metadata,' 'administrative metadata,' 'file groups' and 'structural maps,' the last two of which are intended to group together all of the files that make up a particular digital object and to link content and metadata to a particular structure. The administrative metadata section is intended to store technical information about the file, as well as information about intellectual property rights held in the resource, the source material, and provenance metadata that records relationships between files and migrations. Broadly speaking, the METS schema provides an XML-based container that could be used to store much of the metadata defined in preservation metadata specifications like that published by the Cedars project. Also, a document fully encoded in METS could easily be viewed as an Information Package, as defined by the OAIS model.

NISO draft standard: Technical Metadata for Digital Still Images

As part of a separate initiative, a 'data dictionary' of *Technical Metadata for Digital Still Images* is under review as a draft NISO (National Information Standards Organization) standard (NISO, 2000).

Development of the draft standard first grew out of an "Image Metadata Workshop" held in 1999, sponsored by NISO, the Council for Library and Information Resources (CLIR) and the RLG. The draft standard is not intended to duplicate work on descriptive metadata schemas, but to help define a standardised way of recording the technical attributes of digital images and the production techniques associated with them. The data dictionary includes elements that will record detailed information about images themselves (e.g. formats, compression, etc.), the image creation process, some quality metrics, and any change history (e.g. migrations). No particular encoding of the elements is recommended. Development of the draft standard is based on the experiences of digitisation centres. If and when it is adopted as a standard, it will be of particular use for helping to support the long-term preservation of the products of digital imaging projects.

4.3 *Recordkeeping metadata*

The archives and records professions have also been investigating the metadata that would be required to support the long-term preservation of electronic records. There have been a number of attempts to identify and define recordkeeping metadata; described as "structured or semi-structured information which enables the creation, management, and use of records *through time* and *within* and *across domains* in which they are created" (Hedstrom, 2001, p. 244).

One of the first recordkeeping metadata specifications developed by archivists was developed by the University of Pittsburgh's Functional Requirements for Evidence in Recordkeeping project (Bearman & Sochats, 1996). One significant sign of progress since then has been the development of a general framework known as the Australian Recordkeeping Metadata Schema (RKMS) by a project based at Monash University in Melbourne. The project, amongst other things, attempted to specify and standardise the whole range of recordkeeping metadata that would be required to manage records in digital environments (McKemmish, *et al*., 1999). It has also been concerned with supporting interoperability with more generic metadata standards like the DCMES and relevant resource discovery schemas like the Australian Government Locator Service (AGLS) scheme. The RKMS defines a highly structured set of metadata elements that conforms to a data model based on that developed for the Resource Description Framework (RDF). The schema is designed to be extensible and can inherit metadata elements from other schemas. A group has since been set up to develop the RKMS into an Australian Standard framework for recordkeeping metadata.

In June 2000, a group of archivists, computer scientists and metadata experts met in the Netherlands to discuss metadata developments related to recordkeeping and the long-term preservation of archives (Wallace, 2001). One of the key conclusions made at this working meeting was that the recordkeeping metadata communities should attempt to co-operate more with other metadata initiatives. The meeting also suggested research into the contexts of creation and use, e.g. identifying factors that might encourage or discourage creators from meeting recordkeeping metadata requirements (Hedstrom, 2001, pp. 249-250). This kind of research would also be useful for wider preservation metadata developments. One outcome of this meeting was the setting up of an Archiving Metadata Forum (AMF) to form the focus of future developments.

5 Future Work

Interest in preservation metadata is not limited to the library and recordkeeping sectors but to all organisations and individuals who have an interest in the long-term accessibility or re-usability of digital data. This includes television companies, publishers and other providers of digital content. In the digital library domain, the development of a recommendation on preservation metadata is being co-ordinated by a working group supported by OCLC and the RLG. The membership of the working group is international, and includes key individuals who were involved in the development of the Cedars, NEDLIB and NLA metadata specifications. The key deliverable to date has been a review of the state-of-the-art in preservation

metadata (OCLC/RLG Working Group on Preservation Metadata, 2001a). This includes a summary of the OAIS model, descriptions of the element sets developed by Cedars, NEDLIB and the NLA and an attempt to map between them using the OAIS information model as a general framework. The working group has since published a *Recommendation for Content Information* that provides an expanded conceptual structure for a Content Information package and a set of metadata elements (OCLC/RLG Working Group on Preservation Metadata, 2001b). The recommendation includes elements based on the ones defined in the Cedars, NEDLIB and NLA specifications as well as new elements defined by the working group. Current work includes the production of a recommendation on PDI.

Future work on preservation metadata will need to focus on several key issues. Firstly, there is an urgent need for more practical experience of undertaking digital preservation strategies. Until now, many preservation metadata initiatives have largely been based on theoretical considerations or high-level models like the OAIS. This is not in itself a bad thing, but it is now time to begin to build metadata into the design of working systems that can test the viability of digital preservation strategies in a variety of contexts. This process has already begun in initiatives like the Victorian Electronic Records Strategy (Waugh, *et al*., 2000) and the San Diego Supercomputer Center's 'self-validating knowledge-based archives' (Ludäsher, Marciano & Moore, 2001). A second need is for increased co-operation between the many metadata initiatives that have an interest in digital preservation. This may include the comparison and harmonisation of various metadata specifications, where this is possible. The OCLC/RLG working group is an example of how this has been taken forward within a particular domain. There is a need for additional co-operation with recordkeeping metadata specialists, computing scientists and others in the metadata research community. Thirdly, there is a need for more detailed research into how metadata will interact with different formats, preservation strategies and communities of users. This may include some analysis of what metadata could be automatically extracted as part of the ingest process, an investigation of the role of content creators in metadata provision, and the production of user requirements.

6 Costs

It is very difficult to say anything definite about the costs of creating and maintaining preservation metadata. This is partly a reflection of the current lack of practical experience with such data, but also a recognition that all digital preservation processes involve a commitment in terms of time and money that will be passed on to future generations. The complexity and highly technical nature of preservation metadata suggest that it will be expensive, especially where human intervention in the creation and maintenance processes are required. Chen (2001, p. 26) has said, "the costs incurred in providing and managing adequate metadata will be high."

There may be ways, however, of reducing some of these costs. One way, for example, would be to learn from the experiences of library cataloguing and to try to minimise the duplication of effort through co-operation. One of the main motivations for sharing catalogue records between libraries has been the need to reduce costs. In a digital preservation context, minimising duplication will depend upon timely information being available about which resources digital repositories have attempted to preserve. Also, thought should be given to the development of metadata standards that will permit the easy exchange of preservation metadata (and information packages) between repositories. There is some potentially relevant work currently underway on this in the METS initiative.

A repository might also be able to reduce costs by automating the creation of metadata, wherever this is possible. So, for example, it would be useful if the systems that will need to be developed to facilitate the ingest or migration of digital objects can automatically output metadata about the processes being carried out, and the people and organisations that have authorised them.

As well as ensuring that digital repositories are able to facilitate the automatic capture of metadata, some thought should also be given to how best digital repositories could deal with any metadata that might already exist. This might include, for example, documentation provided by content providers, or technical

parameters recorded as part of a digitisation process. There may be no easy way of ensuring that these are in any standardised format, but the ingest workflow in a digital repository should be able to take account of any metadata that already exists. In the longer term, it may be useful to open a dialogue with the creators and distributors of digital objects concerning the type and form of metadata they create. If they were able to adopt metadata strategies conforming to the best practice for preservation metadata, there would be potential cost-savings for repositories. It is also worth noting that any significant time delay between the creation of a digital object and its ingest into a repository may have adverse cost implications, as there is a possibility that significant information will be lost.

A basic way of reducing costs might be through sharing some metadata within a repository or group of repositories. In practice, it is likely that large numbers of digital objects will require some of the same metadata as they are in a common format or have an identical provenance. It should be possible to design metadata systems that only need to record this shared information once, making it easier to create and update. This approach could be combined with a strategy to migrate all digital objects to an agreed range of common, well-defined formats on ingest or with some kind of canonicalisation (Lynch, 1999). This would help to simplify the technical complexity of objects stored within a digital repository and their preservation strategies, and would have a knock-on effect in that technical metadata would only be required for a relatively limited number of formats. This approach, however, may not be suitable for all types of digital resource.

Generating and maintaining preservation metadata is likely to be expensive but is, however, a prerequisite of ensuring successful digital preservation. The difficulty of preserving digital objects without metadata may mean that it is ultimately a cheaper and more effective option than the alternative. Chen (2001, pp. 26-27) has written that "although more semantics in metadata will increase costs, it will minimise human intervention in accessing data; seamless support, transition of stewardship and lifetime maintenance will improve."

7 Recommendations

7.1 *To Institutions*

Institutions are already acquiring large numbers of digital materials that they will need to preserve. For example, a 1998 survey of RLG member institutions revealed that two-thirds of the respondents owned digital materials for which the institution would need to assume preservation responsibility (Hedstrom & Montgomery, 1998, p. 8). Institutions, therefore, need to help identify these materials, and begin to implement polices that will support their digital preservation, including the generation and maintenance of preservation metadata. Unfortunately, at present there is no set of guidelines that define best practice.

Existing preservation metadata specifications do not tend to specify precisely how they should be implemented, partly because this is dependent upon other considerations, e.g. the system itself. In practice, institutions will need to carefully assess all of their metadata requirements, (e.g. for resource discovery, managing access and preservation), with reference to the published work of existing projects and the recommendations of the OCLC/RLG Preservation Metadata Working Group. This will need to cover the range of resource types that are deemed to need preservation, both online and offline. The assessment may help provide information that could be used to influence the future development of digital preservation strategies and systems. In the short term, the generation and maintenance of metadata may appear expensive, but successful digital preservation strategies depend on it.

In short, institutions should learn to know which of their digital resources require preservation and to devise strategies to ensure their preservation, strategies that will include the identification, generation and maintenance of appropriate metadata.

7.2 *To Creators*

The creators of digital resources are often in the best position to document their technical nature and context. So, for example, publishers of CD-ROMs will know the IPR status and technical operating requirements of any associated software. However, the preservation metadata issue is especially important for projects involved in digitisation. The organisations that fund digitisation programmes are increasingly becoming aware of the potentially short-term nature of their investment. They realise that technical decisions made at the creation stage will have significant effects later on during a digital object's life cycle. For this reason, they are beginning to make funding dependent upon some consideration of digital preservation issues, e.g. in the nof-digitise Technical Standards and Guidelines (UKOLN, 2001).

Digitisation projects have the opportunity to record extremely rich information about the technical nature of digital images, and about the digitisation process itself. The types of preservation metadata that could be recorded during the digitisation process itself include, for example, information about the nature of the source material, the digitisation equipment used and its parameters (formats, compression types, etc.), administrative metadata about the agents responsible for the digitisation process itself, etc. In many cases, the only time that this information can be recorded is as part of the digitisation process itself and, in some cases, it may be possible to generate the output of this metadata automatically from the digitisation software used. The main problem is that there is, as yet, no single standard for this type of metadata. Some guidance on metadata, however, can be found in chapter 5 of the book: *Moving theory into practice* (Kenney & Rieger, 2000).

Preservation metadata may not always just support preservation needs, it may also allow content creators to re-purpose and reuse digital resources in new and interesting ways.

8 Further Reading

Web sites:

PADI: http://www.nla.gov.au/padi/ (visited 1 February 2002).

The main gateway for information on digital preservation is the Preserving Access to Digital Information (PADI) service provided by the National Library of Australia. The service will be found useful by anyone looking for general information about digital preservation issues, but links to specific Web sites relating to preservation metadata can be found at: http://www.nla.gov.au/padi/topics/32.html

Preservation Metadata Working Group: http://www.oclc.org/research/pmwg/ (visited 1 February 2002).

The work of the OCLC/RLG Preservation Metadata Working Group is currently the main international focus on the standardisation of preservation metadata initiatives. The site includes links to a useful state-of-the-art review of preservation metadata, other documents that have been produced by the group, and other resources.

ISO Archiving Standards: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html (visited 1 February 2002).

The OAIS model is becoming more influential in the development of digital preservation systems and the metadata being developed to support them. It is not a particularly easy document to read, but remains the most authoritative source for information on the various models that it defines.

JISC Digital Preservation Focus: http://www.jisc.ac.uk/dner/preservation/(visited 19 October 2001).

Although these Web pages do not focus specifically on preservation metadata, they are a good place to look for information on digital preservation issues in the UK further and higher education sector, and the recently created Digital Preservation Coalition.

RLG DigiNews: http://www.rlg.org/preserv/diginews/ (visited 1 February 2002).

This bimonthly Web newsletter regularly has items on digital preservation and related issues.

Other publications:

Michael Day, Metadata for digital preservation: a review of recent developments. In: P. Constantopoulos and I.T. Sølvberg, eds., *Research and Advanced Technology for Digital Libraries: 5th European*

Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001. (Lecture Notes in Computer Science, 2163). Berlin: Springer, 2001, pp. 161-172. Also available at: http://www.ukoln.ac.uk/metadata/presentations/ecdl2001-day/paper.html (visited 1 February 2002).

This is a recent introduction to preservation metadata that describes some relevant projects and initiatives.

Catherine Grout, Phill Purdy, Janine Rymer, Karla Youngs, Jane Williams, Alan Lock and Dan Brickley, *Creating digital resources for the visual arts: standards and good practice*. Oxford: Oxbow Books, 2000. Also available at: http://vads.ahds.ac.uk/guides/creating_guide/contents.html (visited 1 February 2002).

This is a book in the Guides to Good Practice series produced by the AHDS. This one is produced by the Visual Arts Data Service (VADS) and the Technical Advisory Service for Images (TASi) and covers descriptive metadata, as well as rights management and preservation issues. It notes that "having a digital preservation strategy on board, from the planning stage of a resource, will ensure the longevity and accessibility of the data produced and maximise the investment made in data creation."

Anne R. Kenney and Oya Y. Rieger, *Moving theory into practice: digital imaging for libraries and archives*. Mountain View, Calif.: Research Libraries Group, 2000.

This is probably the best current introduction to digitisation practice for libraries and archives and includes plenty of examples. Chapter 5 covers metadata issues, and is written by Carl Lagoze and Sandra Payette of Cornell University.

Alan Morrison, Michael Popham and Karen Wikander, *Creating and documenting electronic texts*. Oxford: Oxbow books, 2000. Also available at: http://ota.ahds.ac.uk/documents/creating/ (visited 1 February 2002).

This AHDS Guide to Good Practice is produced by the Oxford Text Archive, and covers topics like digitisation, optical-character recognition and SGML and XML-based text markup. Metadata is discussed with regard to the bibliographic headers defined by the Text Encoding Initiative (TEI).

Stuart D. Lee, *Digital imaging: a practical handbook*. London: Library Association Publishing, 2001.

Lee's book is a short introduction to digitisation that includes some useful information on the creation and maintenance of appropriate metadata.

Sean Townsend, Cressida Chappell and Oscar Struijvé, *Digitising history*. Oxford: Oxbow Books, 1999. Also available at: http://hds.essex.ac.uk/g2gp/digitising_history/index.asp (visited 1 February 2002).

This is one of a series of Guides to Good Practice produced by the UK Arts and Humanities Data Service (AHDS) to give advice to the creators of data. This guide, produced by the History Data Service talks about software, data formats, documentation and the importance of preserving historical data in digital form.

9 References

Arms, W.Y., Adkins, R., Ammen, C. & Hayes, A. (2001). Collecting and preserving the Web: the Minerva prototype. *RLG DigiNews*, 5 (2), 15 April. http://www.rlg.org/preserv/diginews/diginews5-2.html (visited 1 February 2002).

Arvidson, A., Persson, K. & Mannerheim, J. (2000). *The Kulturarw3 project: the Royal Swedish Web Archiw3e - an example of "complete" collection of web pages*. 66th IFLA Council and General Conference, Jerusalem, Israel, 13-18 August 2000. http://www.ifla.org/IV/ifla66/papers/154-157e.htm (visited 1 February 2002).

Bearman, D. & Sochats, K. (1996). *Metadata requirements for evidence*. Pittsburgh, Pa.: University of Pittsburgh, School of Information Science. http://www.archimuse.com/papers/nhprc/BACartic.html (visited 1 February 2002).

Chen, S.S. (2001). The paradox of digital preservation. *Computer*, 34 (3), March, 24-28.

Consultative Committee for Space Data Systems. (2001). *Reference model for an Open Archival Information System (OAIS)*. CCSDS 650.0-R-2. Red Book, Issue 2, July 2001. http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html (visited 1 February 2002).

Cunningham, A. (2000). Dynamic descriptions: recent developments in standards for archival description and metadata. *Canadian Journal of Information and Library Science*, 25 (4), 3-17.

Day, M. (1998). *Metadata for preservation*. Cedars project document AIW01. Bath: UKOLN. http://www.ukoln.ac.uk/metadata/cedars/AIW01.html (visited 1 February 2002).

Garrett, J. & Waters, D., eds. (1996). *Preserving digital information: report of the Task Force on Archiving of Digital Information*. Washington, D.C.: Commission on Preservation and Access. Also available at: http://www.rlg.org/ArchTF/ (visited 1 February 2002).

Grout, C. & Ingram, C., eds. (2001). *Working with the Distributed National Electronic Resource (DNER): standards and guidelines to build a national resource*, v. 1.0, February. http://www.jisc.ac.uk/dner/development/guidance/DNERStandards.html (visited 1 February 2002).

Hedstrom, M. (1998). Digital preservation: a time bomb for digital libraries. *Computers and the Humanities*, 31, 189-202.

Hedstrom, M. (2001). Recordkeeping metadata: presenting the results of a working meeting. *Archival Science*, 1 (3), 243-251.

Hedstrom, M. & Montgomery, S. (1998). *Digital preservation needs and requirements in RLG member institutions*. Mountain View, Calif.: Research Libraries Group. http://www.rlg.org/preserv/digpres.html (visited 1 February 2002).

Holdsworth, D. & Sergeant, D.M. (2000). *A blueprint for Representation Information in the OAIS model*. 8th NASA Goddard Space Flight Center Conference on Mass Storage Systems and Technologies and 17th IEEE Symposium on Mass Storage Systems, College Park, Md., USA, 27-30 March 2000. Also available at: http://esdis-it.gsfc.nasa.gov/MSST/conf2000/ (visited 1 February 2002).

Hurley, B.J., Price-Wilkin, J., Proffitt, M. & Besser, H. (1999). *The Making of America II Testbed Project: a digital library service model*. Washington, D.C.: Council on Library and Information Resources. http://www.clir.org/pubs/abstract/pub87abst.html (visited 1 February 2002).

Kenney, A.R. & Rieger, O.Y., (2000). *Moving theory into practice: digital imaging for libraries and archives*. Mountain View, Calif.: Research Libraries Group.

Ludäsher, B., Marciano, R. & Moore, R. (2001). Preservation of digital data with self-validating, self-instantiating knowledge-based archives. *SIGMOD Record*, 30 (3), 54-63.

Lupovici, C. & Masanès, J. (2000). *Metadata for the long term preservation of electronic publications*. NEDLIB report series, 2. The Hague: Koninklijke Bibliotheek. Also available at: http://www.kb.nl/coop/nedlib/results/NEDLIBmetadata.pdf (visited 1 February 2002).

Lynch, C. (1999). Canonicalization: a fundamental tool to facilitate preservation and management of digital information. *D-Lib Magazine*, 5 (9), September. http://www.dlib.org/dlib/september99/09lynch.html (visited 1 February 2002).

Masanès, J. & Lupovici, C. (2001). Preservation metadata: the NEDLIB's proposal. *Zeitschrift für Bibliothekswesen und Bibliographie*, 48 (3-4), 194-199.

McKemmish, S., Acland, G., Ward, N. & Reed, B. (1999). Describing records in context in the continuum: the Australian Recordkeeping Metadata Schema. *Archivaria*, 48, 3-43. Also available at: http://rcrg.dstc.edu.au/publications/archiv01.htm (visited 1 February 2002).

Muir, A. (2001). Legal deposit and preservation of digital publication: a review of research and development activity. *Journal of Documentation*, 57 (5), 652-682.

National Information Standards Organization. (2000). *Data dictionary: technical metadata for digital still images*. Bethesda, Md.: NISO. http://www.niso.org/committees/committee_au.html (visited 1 February 2002).

OCLC/RLG Working Group on Preservation Metadata. (2001a). *Preservation metadata for digital objects: a review of the state of the art*. Dublin, Ohio: OCLC Online Computer Library Center. http://www.oclc.org/research/pmwg/ (visited 1 February 2002).

OCLC/RLG Working Group on Preservation Metadata. (2001b). *A recommendation for content information*. Dublin, Ohio: OCLC Online Computer Library Center, October. http://www.oclc.org/research/pmwg/ (visited 1 February 2002).

Phillips, M., Woodyard, D., Bradley, K., Webb, C. (1999). *Preservation metadata for digital collections: exposure draft*. Canberra: National Library of Australia. http://www.nla.gov.au/preserve/pmeta.html (visited 1 February 2002).

Research Libraries Group. (1998). *RLG Working Group on Preservation Issues of Metadata: final report*. Mountain View, Calif.: Research Libraries Group. http://www.rlg.org/preserv/presmeta.html (visited 1 February 2002).

Ross, S. & Gow, A. (1999). *Digital archaeology: rescuing neglected and damaged data resources*. London: South Bank University, Library Information Technology Centre. Also available at: http://www.hatii.arts.gla.ac.uk/Projects/BrLibrary/ (visited 1 February 2002).

Russell, K. (2000). Digital preservation and the Cedars project experience. *New Review of Academic Librarianship*, 6, 139-154. Also available at: http://www.rlg.org/events/pres-2000/russell.html (visited 1 February 2002).

Russell, K., Sergeant, D., Stone, A., Weinberger, E. & Day, M. (2000). *Metadata for digital preservation: the Cedars project outline specification*. Leeds: Cedars project. http://www.leeds.ac.uk/cedars/metadata.html (visited 1 February 2002).

UKOLN. (2001). *nof-digitise technical standards and guidelines*, v. 3.1, August. http://www.peoplesnetwork.gov.uk/nof/technicalstandards/technicalstandards.html (visited 1 February 2002).

Van der Werf, T. (2000). *The Deposit System for Electronic Publications: a process model*. NEDLIB report series, 6. The Hague: Koninklijke Bibliotheek.

Wallace, D.A. (2001). Archiving Metadata Forum: report from the Recordkeeping Metadata Working Meeting, June 2000. *Archival Science*, 1 (3), 253-269.

Waugh, A., Wilkinson, R., Hills, B. & Dell'oro, J. (2000). Preserving digital information forever. In: *ACM 2000 digital libraries: proceedings of the fifth ACM Conference on Digital Libraries, June 2-7, 2000, San Antonio, Texas*. New York: Association for Computing Machinery, 175-184.

Webb, C. (2000). Towards a preserved national collection of selected Australian digital publications. *New Review of Academic Librarianship*, 6, 179-191.

Appendix A: Structure of the Cedars metadata specification

- Information Package
  - Preservation Description Information
    - Reference Information
      - Resource Description
      - Existing Metadata
        - Existing Records
    - Context Information
      - Related Information Objects
    - Provenance Information
      - History of Origin
        - Reason for Creation
        - Custody History
        - Change History Before Archiving
        - Original Technical Environments
          - Prerequisites
          - Procedures
          - Documentation
        - Reason for Preservation
      - Management Information
        - Ingest Process History
        - Administration History
          - Action History
          - Policy History
      - Rights Management
        - Negotiation History
        - Rights Information
          - Copyright Statement
            - Name of Publisher
            - Date of Publication
            - Place of Publication
            - Rights Warning
            - Contacts or Rights Holder
          - Actors
            - Actions
              - Permitted by Statute
                - Legislation Text Pointer
              - Permitted by License
                - License Text Pointer
    - Fixity Information
      - Authentication Indicator
  - Content Information
    - Representation Information
    - Data Object

Source: Russell, et al*. (2000)*

Appendix B: Abbreviations Used

AGLS	Australian Government Locator Service
AIP	Archival Information Package [OAIS model]
AMF	Archiving Metadata Forum
CCSDS	Consultative Committee for Space Data Systems
Cedars	CURL Exemplars in Digital Archives
CLIR	Council on Library and Information Resources
CPA	Commission on Preservation and Access
CURL	Consortium of University Research Libraries
DCMES	Dublin Core Metadata Element Set
DIP	Dissemination Information Package [OAIS model]
DIS	Draft International Standard
DSEP	Deposit System for Electronic Publications
DNER	Distributed National Electronic Resource
IFLA	International Federation of Library Associations and Institutions
ISO	International Organization for Standardization
JISC	Joint Information Systems Committee
MARC	Machine Readable Cataloguing
METS	Metadata Encoding & Transmission Standard
MOA2	Making of America II project
NEDLIB	Networked European Deposit Library project
NISO	National Information Standards Organization
NLA	National Library of Australia
NOF	New Opportunities Fund
OAIS	Open Archival Information System
OCLC	OCLC Online Computer Library Center, Inc.
PADI	Preserving Access to Digital Information
PANDORA	Preserving and Accessing Networked Documentary Resources of Australia
PDI	Preservation Description Information [OAIS model]
RDF	Resource Description Framework
RKMS	Recordkeeping Metadata Schema
RLG	Research Libraries Group, Inc.
SIP	Submission Information Package [OAIS model]
XML	Extensible Markup Language