ISSN 1747-1524

DCC | Digital Curation Manual

Instalment on

“Preservation Metadata”

http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata

Priscilla Caplan

Digital Library Services

Florida Center for Library Automation

http://www.fcla.edu/

July 2006

Version 1.0

Legal Notices

Catalogue Entry

Title	DCC Digital Curation Manual Instalment on Preservation Metadata
Creator	Priscilla Caplan (author)
Subject	Information Technology; Science; Technology – Philosophy; Computer Science; Digital Preservation; Digital Records; Science and the Humanities.
Description	Instalment on the role of preservation metadata within the digital curation life-cycle. The term is usually reserved for metadata that specifically supports the functions of maintaining the fixity, viability, renderability, understandability, and/or authenticity of digital materials in a preservation context.
Publisher	HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils.
Contributor	Seamus Ross (editor)
Contributor	Michael Day (editor)
Date	14 July 2006 (creation)
Type	Text
Format	Adobe Portable Document Format v.1.3
Resource Identifier	ISSN 1747-1524
Language	English
Rights	© HATII, University of Glasgow

Citation Guidelines

Priscilla Caplan, (July 2006), "Preservation Metadata", *DCC Digital Curation Manual*, S.Ross, M.Day (eds), Retrieved <date>, from http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata

About the DCC

The JISC-funded Digital Curation Centre (DCC) provides a focus on research into digital curation expertise and best practice for the storage, management and preservation of digital information to enable its use and re-use over time. The project represents a collaboration between the University of Edinburgh, the University of Glasgow through HATII, UKOLN at the University of Bath, and the Council of the Central Laboratory of the Research Councils (CCLRC). The DCC relies heavily on active participation and feedback from all stakeholder communities. For more information, please visit www.dcc.ac.uk. The DCC is not itself a data repository, nor does it attempt to impose policies and practices of one branch of scholarship upon another. Rather, based on insight from a vibrant research programme that addresses wider issues of data curation and long-term preservation, it will develop and offer programmes of outreach and practical services to assist those who face digital curation challenges. It also seeks to complement and contribute towards the efforts of related organisations, rather than duplicate services.

DCC - Digital Curation Manual

Editors

Peer Review Board

Preface

The Digital Curation Centre (DCC) develops and shares expertise in digital curation and makes accessible best practices in the creation, management, and preservation of digital information to enable its use and re-use over time. Among its key objectives is the development and maintenance of a world-class digital curation manual. The *DCC Digital Curation Manual* is a community-driven resource—from the selection of topics for inclusion through to peer review. The Manual is accessible from the DCC web site (http://www.dcc.ac.uk/resource/curation-manual).

Each of the sections of the *DCC Digital Curation Manual* has been designed for use in conjunction with *DCC Briefing Papers*. The briefing papers offer a high-level introduction to a specific topic; they are intended for use by senior managers. The *DCC Digital Curation Manual* instalments provide detailed and practical information aimed at digital curation practitioners. They are designed to assist data creators, curators and re-users to better understand and address the challenges they face and to fulfil the roles they play in creating, managing, and preserving digital information over time. Each instalment will place the topic on which it is focused in the context of digital curation by providing an introduction to the subject, case studies, and guidelines for best practice(s). A full list of areas that the curation manual aims to cover can be found at the DCC web site (http://www.dcc.ac.uk/resource/curation-manual/chapters). To ensure that this manual reflects new developments, discoveries, and emerging practices authors will have a chance to update their contributions annually. Initially, we anticipate that the manual will be composed of forty instalments, but as new topics emerge and older topics require more detailed coverage more might be added to the work.

To ensure that the Manual is of the highest quality, the DCC has assembled a peer review panel including a wide range of international experts in the field of digital curation to review each of its instalments and to identify newer areas that should be covered. The current membership of the Peer Review Panel is provided at the beginning of this document.

The DCC actively seeks suggestions for new topics and suggestions or feedback on completed Curation Manual instalments. Both may be sent to the editors of the *DCC Digital Curation Manual* at curation.manual@dcc.ac.uk.

Seamus Ross & Michael Day.
18 April 2005

Biography

Priscilla Caplan is Assistant Director for Digital Library Services at the Florida Center for Library Automation. She has been involved with digital preservation since 2001, when she began planning the development of the FCLA Digital Archive, a preservation repository for the use of the public universities of Florida. With Rebecca Guenther, she co-chaired the OCLC/RLG Working Group on Preservation Metadata: Implementation Strategies (PREMIS) and continues to be involved with the PREMIS Maintenance Activity. She is the author of *Metadata Fundamentals for All Librarians* (ALA Editions, 2003).

Table of Contents

Introduction and scope

Preservation metadata is information that supports and documents the process of digital preservation. The term is usually reserved for metadata that specifically supports the functions of maintaining the fixity, viability, renderability, understandability, and/or authenticity of digital materials in a preservation context. As such, preservation metadata includes elements of administrative metadata, structural metadata, and technical metadata (the subset of administrative metadata that documents detailed format characteristics of files). It can also include some rights metadata – the documentation of intellectual property rights, permissions, and restrictions on use. Descriptive metadata that primarily supports discovery and access is not generally considered preservation metadata, although it is certainly necessary information for most preservation repositories.

A recent Digital Preservation Coalition *Technology Watch Report* gives an excellent overview of preservation metadata.[1]

Background and developments to date

Libraries and archives have taken different approaches to preservation metadata. For libraries, the evolution of a metadata framework began with the seminal 1996 report *Preserving Digital Information.*[2] This report stated

Whatever preservation method is applied ... the central goal must be to preserve information integrity; that is, to define and preserve those features of an information object that distinguish it as a whole and singular work. In the digital environment, the features that determine information integrity and deserve special attention for archival purposes include the following: content, fixity, reference, provenance, and context.

It went on to elaborate the features of each of these five attributes of information integrity. "Content" referred to the object to be preserved, although the report recognized that identifying content was not always straightforward. A key insight was that certain preservation strategies, for example, format migration, require that an object be changed in order to be preserved. Content therefore might not always be defined as a particular set of bits but may have to be abstracted to qualities of structure and format, or even to abstract intellectual meaning. "Fixity" referred to the way that content was fixed as a discrete object, and mechanisms for preventing or detecting change. "Reference" referred to means of identifying, citing and locating digital works. "Provenance" meant the record of the origin and chain of custody of the digital object. "Context" was defined rather broadly as the ways in which digital objects "interact with elements in the wider digital environment." It included hardware, software and media dependencies, as well as linkages among digital objects and even "social context."

The drafters of the Open Archival Information Systems Reference Model (OAIS) moved these attributes into a metadata context when they used the same categories in the OAIS information model. [3] The OAIS is a framework for understanding the requirements of long-term preservation systems, defining both a functional model and an information model for preservation activities.[4] Because the OAIS information model

[1] Lavoie, Brian and Richard Gartner, September 2005, *Technology Watch Report: Preservation Metadata*, http://www.dpconline.org/docs/reports/dpctw05-01.pdf [Accessed: 22 November 2005, 12:23].

[2] Waters, Donald and John Garrett, 1996, *Preserving Digital Information: Final Report of the Task Force on Archiving of Digital Information*, ftp://ftp.rlg.org/pub/archtf/final-report.pdf [Accessed: 22 November 2005, 12:51].

[3] Consultative Committee for Space Data Systems, January 2002, *Reference Model for an Open Archival Information System (OAIS)*, http://www.ccsds.org/documents/650x0b1.pdf [Accessed 22 November 2005, 12:59].

[4] Note that the term "OAIS" is used variously to refer to the *Reference Model for an Open Archival Information System* publication, the reference model itself, and a repository conforming to the requirements of the reference model.

has influenced nearly all subsequent work in preservation metadata, it is worth describing in some detail.

In the OAIS model an Information Package consists of Content Information and Preservation Description Information, held together by Packaging Information.[5] Content Information includes both the Content Data Object to be preserved (the *ding an sich*, as it were) and Representation Information, data that make the Content Data Object understandable. For example, if the Content Data Object were a numeric dataset, its Representation Information might include documentation explaining the format of each record (structure) and the meaning of each numeric variable (semantics). The Representation Information might in turn require more Representation Information, forming a "chain" of Representation Information. In OAIS, both the Content Data Object and its Representation Information must be treated equally as the target of preservation. In practice, some Representation Information might be a Content Data Object in its own right (e.g. a codebook or format specification) while other Representation Information might be what we would think of as metadata (e.g. MIME type or number of bits per sample).

Preservation Description Information is the information necessary to preserve the Content Information. It consists of information expressing aspects of the attributes of reference, context, provenance and fixity, taken directly from the 1996 report. Reference Information identifies mechanisms used to assign identifiers to Content Information. Context Information "documents the relationships of the Content Information to its environment" and includes information about why the Content Information was created as well as its relationship to other Content Information.[6] Provenance Information documents the creation, modification, and custody of the Content Information. Fixity Information describes the checks or keys such as checksums used to ensure that the Content Information has not been changed in an unauthorized manner.

The OAIS information model defines these categories of information and gives loose examples of the types of metadata that might fall within each category. Reference Information, for example, might include an object identifier while Fixity Information might include a checksum. The model does not define any specific metadata elements, and focuses more on the uses and transformations of various types of Information Packages than on the detailed contents of the package itself. The OAIS was first published in draft form in 1999, and it is credited with influencing most subsequently published preservation metadata specifications, which generally include elements of both Representation Information and of Preservation Description Information.[7] In 2003 it was approved as ISO Standard 14721.

The National Library of Australia (NLA) was one of the first institutions to actually build a digital archive with the establishment of the PANDORA archive of web-accessible materials in 1996. The NLA has been a leader in developing a collaborative national approach to the long-term preservation of Australian publications, understanding that "archiving Australian online publications is only the first step in ensuring long-

[5] Capitalized terms, such as "Information Package" and "Content Information," are terms capitalized and defined in the *Reference Model for an Open Archival Information System*.

[6] In a small deviation from *Preserving Digital Information*, it does not include media dependencies, which are considered part of packaging rather than Preservation Descriptive Information.

[7] An influential specification by the RLG Working Group on the Preservation Uses of Metadata, however, predated the OAIS. By 1997 libraries microfilming for preservation had just begun to incorporate digitization as well, and the Working Group was established to identify the metadata that should be recorded for digital masters. Their final report, which became known as the PRESERV specification, defined 16 metadata categories geared specifically to digital images created by scanning or photographing non-digital originals, including such aspects as capture device, compression and color management.

term access to them."[8] In 1999 the NLA published its "Preservation Metadata for Digital Collections" for public comment.[9] The document claimed to be "informed" by the OAIS model, as well as by the PANDORA experience and by other experiments in digital archiving.

The NLA document was a significant attempt to move from theory towards practice. It defined twenty high-level metadata elements with some specificity, including attributes such as repeatability and obligation. In addition, it recognized the need to associate metadata with a data model, and distinguished metadata elements appropriate to collections, objects, and sub-objects (files). It recognized that different file formats required different elements of technical metadata and defined different lists of "file description" subelements for image, audio, video, text, database and executable files.

Two other influential specifications, those of the CEDARS and NEDLIB projects, were released in 2000. CEDARS (CURL Exemplars in Digital Archiving) was a project of the U.K. Consortium of University Research Libraries to explore strategic, methodological and practical issues in digital preservation. The CEDARS metadata specification explicitly attempted to translate the abstract OAIS model into more practical metadata specifications, albeit in the context of a research project.[10] CEDARS defined preservation metadata broadly as the information required "to support meaningful access to the archived digital content and includes descriptive, administrative, technical and legal information."[11]

The Networked European Deposit Library (NEDLIB) was a collaborative project of European national libraries led by the National Library of the Netherlands. NEDLIB's metadata specification was also explicitly based on OAIS. Unlike CEDARS, however, NEDLIB focused specifically on the metadata needed to address problems of technical obsolescence. According to NEDLIB, "The information involved in long term preservation metadata is information about the data processing of the digital objects that are to be preserved... The problem is to describe precisely this processing or/and the elements involved in it, because the modality of data processing will be different in 10 or 200 years. That is why the image format (unlike its resolution) is part of the preservation information."[12]

The same year as the NEDLIB and CEDARS papers were released, OCLC and RLG established a joint Working Group on Preservation Metadata (posthumously renamed the Preservation Metadata Framework Working Group). The first report of the group, "Preservation Metadata for Digital Objects: A Review of the State of the Art," compared and contrasted the NLA, CEDARS and NEDLIB specifications in terms of their rationales and objectives, their underlying frameworks, and their defined metadata elements, using the OAIS information model to organize the elements for comparison.[13] This interesting analysis illuminated the OAIS model as well as the competing metadata specifications.

This was followed in 2001 by the Working Group's second report, "Preservation Metadata

[8] http://pandora.nla.gov.au/key_docs.html [Accessed 22 November 2005, 13:54].

[9] National Library of Australia, 15 October 1999, *Preservation Metadata for Digital Collections*, http://www.nla.gov.au/preserve/pmeta.html [Accessed 22 November 2005, 13:56].

[10] *Metadata for Digital Preservation: The CEDARS Project Outline Specification Draft for Public Consultation*, March 2000, http://www.leeds.ac.uk/cedars/cedars.pdf [Accessed 22 November 2005, 13:59].

[11] ibid, p. 1.

[12] Lupovici, Catherine and Julien Masanès, July 2000, *Metadata for Long-term Preservation*, http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm [Accessed 22 November 2005, 14:05].

[13] OCLC/RLG Working Group on Preservation Metadata, 31 January 2000, *Preservation Metadata for Digital Objects: A Review of the State of the Art*, http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf [Accessed 22 November 2005, 14:08].

and the OAIS Information model: A Metadata Framework to Support the Preservation of Digital Objects."[14] The Framework elaborated and in some places expanded the OAIS Information model, and like the earlier report used this structure to organize a set of metadata elements. It defined a synthesis (that is, a de-duplicated superset) of elements from NLA, CEDARS, NEDLIB and a fourth scheme used by the OCLC Digital Archive. The synthesis was then supplemented by "refinements, elaborations, and additional structure and elements recommended by the working group members."[15] As noted by Michael Day, the Framework effectively superseded the specifications it was based upon and represented "a good starting point for future practical implementations of preservation metadata."[16]

Even so, the Framework did not define a metadata scheme that could be used in practice by a preservation repository. It had no underlying data model, leading to some ambiguity as to the nature of objects being described. Also, most elements were defined in such a way as to discourage manipulation by computer, some appearing to require lengthy narrative descriptions as values. The National Library of New Zealand, finding past work too theoretical, developed its own preservation metadata element set in 2002 and 2003.[17] Metadata elements were adapted from earlier specifications but defined more rigorously, with some attempt to maximize the potential for automation. The specification defined a data model with four types of entities: objects, processes, files and metadata modification. The last showed the recognition that the metadata record itself was important data that must be secured and managed over time.[18]

In 2003 OCLC and RLG established a second international working group to take the analysis of the Preservation Metadata Framework group to the next step, and develop an implementable core set of preservation metadata elements generically applicable to preservation repositories. Called PREMIS (PREservation Metadata: Implementation Strategies), the group was composed mostly of representatives of institutions developing or operating preservation repositories. Taking a highly pragmatic approach, they defined core metadata as "the things that most working preservation repositories are likely to need to know in order to support digital preservation."[19] The PREMIS Data Dictionary was issued in 2005 and awarded the Digital Preservation Coalition's Digital Preservation Award for that year.

The PREMIS data model defines five types of entity: Intellectual Entities (that is, conceptual objects that might be composed of one or more digital files), Objects, Rights, Agents and Events. The PREMIS Data Dictionary is organized around these types of entities rather than the categories of the OAIS information model, but they can be mapped to each other. Metadata pertaining to Objects includes what the OAIS information model would call Reference Information (identifiers), Fixity Information (message digests and digital signatures), some Context Information (relationships and environment), and Representation Information (object

[14] OCLC/RLG Working Group on Preservation Metadata, June 2002, *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*, http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf [Accessed 22 November 2005, 14:10].

[15] ibid, p. 3.

[16] Day, Michael, *Preservation metadata. Prepublication draft of chapter published in: G. E. Gorman and Daniel G. Dorner (eds.), Metadata applications and management*, I*nternational Yearbook of Library and Information Management, 2003-2004, London: Facet Publishing, 2004, pp. 253-273,* http://www.ukoln.ac.uk/metadata/publications/iylim-2003/ [Accessed 22 November 2005, 14:13].

[17] National Library of New Zealand, July 2003, *Metadata Standards Framework – Metadata Implementation Schema*, http://www.natlib.govt.nz/files/nlnz_data_model.pdf [Accessed 22 November 2005, 14:15].

[18] Searl, Sam and Dave Thompson, *Preservation Metadata: Pragmatic First Steps at the National Library of New Zealand*, http://www.dlib.org/dlib/april03/thompson/04thompson.html [Accessed 22 November 2005, 14:18].

[19] PREMIS Working Group, May 2005, *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*, http://www.oclc.org/research/projects/pmwg/premis-final.pdf [Accessed 22 November 14:24].

characteristics.[20] OAIS Provenance Information and some Context Information is expressed through metadata pertaining to Agents and Events.

PREMIS is notable for insisting on a certain amount of rigor in the application of preservation metadata. It defines three different types of Object – files, bitstreams within files, and representations (the set of all files needed to render an Intellectual Entity) – and requires metadata creators to distinguish between them. A web page as rendered, for example, might consist of an HTML file and dozens of gif images. PREMIS recognizes the importance of recording both metadata pertaining to the web page as a whole and metadata pertaining to each HTML and gif file.

PREMIS work is continuing under the rubric of the PREMIS Maintenance Activity, supported by OCLC, RLG, the Florida Center for Library Automation, and the U.S. Library of Congress. The Library of Congress has committed some funds to the development of more documentation and to further exploration of how PREMIS metadata might best be obtained. Plans for training events and successor initiatives are under consideration. There is a small, appointed Editorial Board, and a larger PREMIS Implementers' Group (PIG) open to anyone actively engaged in using the PREMIS Data Dictionary. The PIG has a discussion list and a wiki (the PIG Pen) where members are encouraged to post examples of their use of PREMIS metadata. XML schemas to support exchange of PREMIS metadata are available on the Maintenance Activity website.[21]

Although archivists had some representation on the PREMIS Working Group and are encouraged to participate in the PIG, the archival approach to preservation and to preservation metadata has followed a different course which can be traced back to the Pittsburgh Project in the mid-1990s. Formally titled "Functional Requirements for Evidence in Recordkeeping," this research project was led by Richard Cox and David Bearman at the University of Pittsburgh School of Information Sciences in the U.S. It attempted to define fundamental properties of records and functional requirements for recordkeeping for use in the design of electronic information systems. Bearman saw records as primarily evidentiary, and introduced the concept of metadata for recordkeeping, focusing heavily on the property of records as evidence of business transactions. The project published a metadata specification based on these ideas entitled "A Reference Model for Business Acceptable Communications" (BAC).[22]

Conterminously with the Pittsburgh Project, the University of British Columbia (UBC) in Canada carried out a research project to identify requirements for creating, handling and preserving reliable and authentic electronic records. The project focused on means to ensure the authenticity and reliability of electronic records and on management issues related to the maintenance and preservation of reliable and authentic records. Although the methodology and overall approach differed from that of the Pittsburgh Project, UBC also emphasized the evidentiary value of records in their focus on reliability, the authority of a record as evidence. The project created a set of eight templates (metadata) for identifying essential components of records in various recordkeeping systems.

[20] The OAIS concept of Representation Information (information needed to ensure an object remains understandable) should not be confused with the PREMIS concept of a Representation (the set of all files and information needed to render a usable version of an Intellectual Entity).

[21] http://www.loc.gov/standards/premis/ [Accessed 22 November 2005, 14:31].

[22] Metadata Specifications Derived from the Fundamental Requirements: A Reference Model for Business Acceptable Communications, http://web.archive.org/web/20000302194819/www.sis.pitt.edu/~nhprc/meta96.html [Accessed 14 April 2006, 18:38]. Somewhat ironically, the original website for the Pittsburgh Project including the BAC itself was destroyed, but it can still be accessed via the Internet Archive's Wayback Machine.

Both projects began streams of influence that continue to the present day. According to Michael Day, "The Pittsburgh Project inspired the development of a whole new series of recordkeeping metadata initiatives, especially in Australia," including the *Recordkeeping Metadata Standard* of the National Archives of Australia and the Victorian Electronic Records Strategy (VERS) initiative of the Public Record Office Victoria.[23] Similarly, the UBC project spawned InterPARES (International Research on Permanent Authentic Records in Electronic Systems), a two-phase international collaboration also led by the University of British Columbia. InterPARES 1 (1999-2001) addressed the selection and preservation of authentic electronic records generated in databases and document management systems used in administrative and legal activities. InterPARES 2 (2002-2006) is concerned with records produced in new "experiential, dynamic, and interactive" digital environments resulting from artistic, scientific and government activities.[24] Although neither phase focused on metadata specifically, both had metadata components, and InterPARES 2 developed a metadata schema registry to help assess the recordkeeping and archival capabilities of existing metadata schemes.

The long-term focus of the archival community on recordkeeping has led to several differences between their approach and that of librarians to preservation and preservation metadata. The early insistence of the Pittsburgh Project that records are valuable only insofar as they represent business transactions and that the aim of digital archives is to preserve their evidentiary value has been as influential as it has been controversial. It forced attention onto the records management, as opposed to curatorial, functions of archivists, and deprecated the importance of records in conveying information, documenting history, and preserving memory.[25] There are several apparent consequences. First, archivists tend to see digital preservation more as a component of an electronic recordkeeping system than as a discrete function in its own right; they are less likely than librarians to think in terms of preservation repositories. Similarly, preservation metadata is rarely considered alone but is rather an integral and inextricable part of recordkeeping metadata.

It also follows that archivists are currently more willing than librarians to discard the original form, look and feel of a source document so long as the content and evidentiary value is preserved and the authenticity of the content can be demonstrated. In this view preservation metadata must define the essential characteristics of a record, and will vary according to the situation. If the content is essential, then the markup of that content can be considered preservation metadata. If the font is also essential, then preservation metadata must record the font.

Preservation metadata elements

Preservation metadata represents a repository's best guess as to what information will be necessary in order to make it possible to use a digital item in the future, given the likelihood of changes in technology, format obsolescence, and other risks. The nature of the projected use may vary depending on the nature of the item, the user community for which it is being preserved, and the institution responsible for its preservation. Different preservation strategies may also demand that different pieces of information be recorded. For these reasons there is no universal preservation metadata element set and no expectation that there will or should ever be one. Even PREMIS attempts only to be a core set of

[23] Day, Michael, *Preservation Metadata*, p. 265.

[24] InterPARES 2 Project home page, http://www.interpares.org/ [Accessed 14 April 2006, 19:17].

[25] See Rosenzweig, Roy, June 2003, "Scarcity or Abundance? Preserving the Past in a Digital Era," in *The American Historical Review*, http://www.historycooperative.org/journals/ahr/108.3/rosenzweig.html [Accessed 15 April 2006, 9:42].

"things that most working preservation repositories are likely to need to know in order to support digital preservation," with the words "most" and "likely" carefully chosen to allow wiggle room.[26] With this understanding, some examples of commonly used preservation metadata elements are given below.

Format identification
Obviously, it is important to record the format of a digital file, although this is not nearly as straightforward as it might appear at first glance. Format designations in common use, such as file extensions and MIME types, are insufficiently granular and do not distinguish between versions. Format registries, such as the U.K. National Archives' PRONOM and the proposed Global Digital Format Registry aim to address this problem by assigning a unique identifier to each format. There is also some judgment involved in what constitutes a format. Depending on the repository, an SGML document conforming to the DocBook DTD could be considered to be in text, SGML, or DocBook format. To complicate things further, the codec (compression/decompression) used within a file may be as important as the format; for example, the video stream in an AVI file might be coded in CinePak, Motion JPEG or Indeo.

Along with the format itself, preservation metadata might record where a file fails to conform to the format specification (anomalies) and whether encryption or other devices are employed to restrict use (inhibitors).

Significant properties
Significant properties are characteristics of an item that should be preserved through future migrations or emulation. The determination of significant properties may be a repository-wide decision adhering to all materials in a particular class (for example, a policy that only the textual content of memos must be preserved) or it may be specific to particular items. The Arts and Humanities Data Service, for example, commits to defining the significant properties of each deposited digital resource in consultation with the depositor.[27]

Environment for use
Records of the hardware, software and ancillary files required to render or use a digital object are collectively known as environment information. This is important preservation metadata, as most preservation strategies require knowledge of the larger environment. In the case of a database, for example, the data tables themselves are of little use without the database model or schema. If the database was exported in the proprietary format of the database management system, a copy of the same database management system, possibly even to the same release version and operating system, might be required for access. Environment information can get complicated fast. Even in the relatively simple case where only a browser is needed to render an item, the browser may run only on certain microcomputer operating systems, which in turn may run only on particular models of computer, and only when meeting minimum requirements for processor speed and memory. Happily, environment data can often be associated with the file format, making it possible to consider the development of shared registries of environment information.

Fixity
Fixity information is essential for determining whether a file has been changed between two points in time. Most commonly, a message digest (informally called a "checksum") is computed by applying a hashing algorithm over the content of the file. If repeating the process at a later time

[26] PREMIS Working Group, p. ix.

[27] James, Hamish, 2004, *Collections Preservation Policy*, http://ahds.ac.uk/documents/colls-policy-preservation-v1.pdf [Accessed 15 April 2006, 13:34].

produces a different message digest, the file has been altered. At a minimum, metadata must record the hashing algorithm used and the message digest produced for future comparison.

Technical metadata

Much preservation metadata is "technical metadata," or metadata describing the technical properties of digital files and bitstreams. Some of these properties, such as size, format, and fixity information, are applicable to most materials and are included in PREMIS and other general preservation metadata specifications. Other properties are specific to particular file types and/or formats. Bit depth, for example, pertains to audio and image files, while character encoding pertains to text files. There is room for debate as to what particular characteristics are important to record for various file types, and few agreed upon standards.

The specification of detailed technical metadata for images is the most advanced. The National Information Standards Organization (NISO) draft standard *Data Dictionary - Technical Metadata for Digital Still Images* (Z39.87) defines metadata pertaining primarily to scanned images such as TIFF preservation masters.[28] There is an XML representation of the data dictionary called MIX (Metadata for Images in XML Schema) maintained by the Library of Congress.[29]

Technical metadata requirements for other formats are less well developed. The National Library of Australia's "Preservation Metadata for Digital Collections" contain basic technical metadata elements specific to image, audio, video, text, database and executable files. Technical metadata specifications defined by the Library of Congress' Audio-Video Prototyping project for audio, image, text and video content have been used by a number of preservation projects but have not led to any standardization effort.[30] The Audio Engineering Society has a draft standard for Core Audio Metadata that has also been used as the basis for local preservation metadata schemes.

Provenance

Digital provenance documents the origin and chain of custody of a digital object, and any important events in the object's history. Some repositories consider digital provenance to include a history of changes since an object's creation, such as migrations, normalizations and other transformations. In the PREMIS model, an object cannot be changed; the act of modification creates a new object related to the source by derivation. In any case, metadata pertaining to digital provenance includes information about an object's creators, owners and rights holders, as well as a record of actions (events or processes) affecting the object from the time of its creation. The PREMIS data dictionary includes semantic units describing events, such as the nature of the event, the date and time it occurred, the object(s) and agent(s) involved, and the outcome. The preservation metadata defined by the National Library of New Zealand includes elements describing processes, which are conceptually similar to PREMIS events but may involve multiple steps over an extended period of time. New Zealand also records the reason why the process was carried out.

Packaging

Another aspect of preservation metadata is the packaging needed to bundle metadata together with content. Here the Metadata Encoding and Transmission Standard (METS) has become a *de*

[28] *Data Dictionary - Technical Metadata for Digital Still Images*, http://www.niso.org/standards/resources/Z39-87-200x-forballot.pdf?CFID=6860130&CFTOKEN=81464797 [Accessed 22 November 2005, 14:50].

[29] *NISO Metadata for Images in XML Schema Official Web Page*, http://www.loc.gov/standards/mix/ [Accessed 22 November 2005, 14:54].

[30] http://www.lcweb.loc.gov/rr/mopic/avprot/extension2.html [Accessed 22 November 2005, 15:01].

facto standard for libraries and archives.[31] METS is an XML schema that defines the hierarchical structure of a digital object and relates that structure to a list of all files included in the object. The files themselves can be linked to or embedded within the METS document. Additional metadata can be supplied by the use of "extension schema," a convenient way to plug in descriptive or administrative metadata created according to an independent metadata schema. METS is commonly used in preservation repository applications for Submission Information Packages (SIP) and Dissemination Information Packages (DIP), types of information packages defined by OAIS.

Other content packaging standards are dominant within other communities. The IMS Content Packaging Specification (IMS-CP) developed by the IMS Global Learning Consortium is used in education.[32] The MPEG-21 Digital Item Declaration Language (DIDL) is used in commercial applications and has some proponents in the digital library community.[33] The Consultative Committee for Space Data Systems (the group who developed the OAIS reference model) is developing a packaging specification called XFDU, the XML Formatted Data Unit.[34] All of these are XML schema with the potential for use in the preservation environment.

Application to Digital Curation

The question of who should create preservation metadata is not as simple as it might first appear. Because of the great number of even "core" preservation metadata elements and the degree of accuracy and consistency required, it is widely believed that metadata values should be supplied as much as possible automatically by software applications. Quite a bit of technical information about an object can be obtained automatically by parsing the object and extracting or inferring object characteristics. For example, the National Library of New Zealand has developed an open source metadata extraction tool that automatically extracts preservation-related metadata from digital files and outputs it in an XML format for loading into other systems.[35] Other information, such as the hardware and software environments required to render an object in a particular format, can be made available in central registries for look-up by program. Once the object is in the repository, information about its storage location and about some of the actions performed on it (events) can also be supplied by the programs that store and manipulate it. For example, a program that performs a fixity check on a file by calculating a message digest and comparing it to an earlier one can record that check as an event.

Data Creators

Within the Digital Curation Centre constituency, data creators are the scientists and scholars who produce digital data on a regular basis. It is widely accepted that publicly funded research data is a public good and that it should be made available for sharing and reuse. Data creators are obliged to keep future usability as well as current use in mind, and to take steps to help others find, use, and curate the data. According to a National Science Foundation report on long-lived data collections, data authors have among their responsibilities to:

- conform to community standards for recording data and metadata that

[31] METS Metadata Encoding and Transmission Standard Official Website, http://www.loc.gov/standards/mets/ [Accessed 22 November 2005, 15:07].

[32] http://www.imsglobal.org/content/packaging/ [Accessed 30 November 2005, 10:25].

[33] Bekaert, Jeroen, Patrick Hochstenbach and Herbert Van de Sompel, November 2003, *Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library*, D-Lib Magazine v.9:no.11, http://www.dlib.org/dlib/november03/bekaert/11bekaert.html [Accessed 30 November 2005, 2:15].

[34] See http://sindbad.gsfc.nasa.gov/xfdu/ [Accessed 30 November 2005, 17:29].

[35] http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction [Accessed 15 April 2006, 10:17].

adequately describe the context and quality of the data and help others find and use the data;

- conform to community standards for the type, quality and content of data including associated metadata, for deposition in relevant data collections;
- develop and continuously refine a data management plan that describes the intended duration and migration path of the data.[36]

A key role of the data creator is in creating and/or providing any necessary Representation Information required to make the archived object understandable. The data creators may be the only parties with a complete understanding of the structure and semantics of the data. For example, if the material to be archived is a social science dataset, Representation Information might include syntax statements defining the raw data file, the codebook, and the data collection instrument. For science and social science data, documentation should provide information about the methodology and procedures used to collect the data, details about codes, definitions of variables, variable field locations, and the like. This is needed equally for near-term reuse and long-term preservation of the dataset.

The data creator is not in general expected to be the key creator of preservation metadata such as format-specific details or hardware and software environment. Much of this information will be obtained if possible by automatic methods mentioned above. In practice, documentation such as a codebook will be treated by a preservation repository as another Content Data Object to be archived. Preservation metadata pertaining to the codebook as a digital object will be extracted or otherwise supplied by the same methods that supply metadata for the dataset itself.

In some areas, however, the input of the data creator can be critical. One of these is the determination of "significant properties," those aspects of the object that are significant for preservation. The content, behavior, functionality, structure and appearance of the original object may or may not be essential characteristics. In a textual document, for example, preservation of bolding and other typographic conventions may be necessary to convey emphasis, while hotlinks between sections of the document are expendable conveniences. For a database, the determination of significant properties can be quite complex, as the utility of the database may depend in part upon the data itself, in part on procedures executed by the database management system, and in part on the application invoking the database management system.[37] Data creators, data curators and data users may all need to play a part in agreeing to the definition of significant properties for any given object.

Data creators are also the most likely source of information about events that occur in the lifecycle of the object before it is ingested into the preservation repository. This is an area that receives little attention in the OAIS model and (possibly as a result) is largely neglected in PREMIS. Information about the creation, maintenance and change history of material before it is archived is relevant to its quality, authenticity and provenance and often can only be supplied by the data creator. Pre-ingest activities such as negotiations between the data creator and the archive, the transfer of data to the archive, and validation processing and follow-up are probably best recorded by the repository.[38]

[36] National Science Board, October 2005, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf [Accessed 22 November 2005, 15:20].

[37] Verdegem, Remco, *Database Preservation Issues*, http://www.digitaleduurzaamheid.nl/bibliotheek/docs/longterm_preservation_of_databases.pdf [Accessed 22 November 2005, 15:28].

[38] Beedham, Hilary, Matt Palmer, and Raivo Ruusalepp, *Assessment of UKDA and TNA Compliance with OAIS and METS Standards*,

Finally, data creators may be the authoritative source of information about intellectual property rights affecting preservation. The data creator(s) may individually or jointly be the rights holder, or rights may belong to the creators' institution(s), to the funding organization, or to the public.

Data Curators

Data curators manage data to ensure its availability for discovery and re-use. For long-lived data collections, those expected to be needed for a period of time long enough for there to be concern about the impacts of changing technology, the curator will be concerned not only with near-term sustainability, but also with long-term preservation. Data managers have the responsibility to "provide for the integrity, reliability and preservation of the collection by developing and implementing plans for backup, migration, maintenance and all aspects of change control."[39]

The data curator is responsible for preservation metadata as part of his overall responsibility for preservation. Whether the organization is developing its own preservation system, implementing a third-party preservation repository, or outsourcing preservation functions to another organization, the curator is responsible for ensuring that the preservation metadata recorded is adequate for the goals of the repository.

The data curator has a particularly important role in establishing trust. It is important to preserve an assessment of the quality of the data and supporting information concerning data cleaning, data validation, and known problems. This metadata is increasingly important over time. As noted by a 2003 report on e-Science Curation, "data we should now doubt may in the future be assumed to be correct."[40]

The same report points out that the usefulness of data may depend on tools for access, visualization, manipulation, etc., which may themselves require preservation action. The curator, working with the data creators, must document these dependencies. Preservation metadata includes information about software environments required to render and use the archived materials as well as the hardware environments supporting such software. As noted above, this can be extremely difficult to supply in detail, leading to speculation that centrally maintained registries are the best way to obtain such information. However, the more specialized the tool, the less likely it is to be documented in central registries.

Data Users

Data Users play a critical part in the OAIS model, because they constitute the Designated Community. In OAIS, the Designated Community is the set of consumers who should be able to understand the preserved information, and as such, the knowledge base of the Designated Community determines the minimum amount of Representation Information that an OAIS must preserve.

In practice, this means that data curators must understand the knowledge, skills and point of reference of the data users, and distill from this the baseline of information which the majority of data users can be expected to know without consulting the experts who produced the data. In many cases data curators will have this understanding because they themselves are part of the Designated Community, or because they work closely with data users to provide access and

http://www.data-archive.ac.uk/news/publications/oaismets.pdf [Accessed 22 November 2005, 15:32].

[39] National Science Board, October 2005, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf, p. 26 [Accessed 22 November 2005, 15:20].

[40] Lord, Philip, and Alison Macdonald, 2003, *Data Curation for e-Science in the UK: An Audit to Establish Requirements for Future Curation and Provision*, http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf, p. 53.

reference services. Data users must always be willing to work with data curators to ensure the curators' understanding is accurate and keeps up with changing conditions.

Preservation metadata in action

To the extent that any repository performs preservation functions, it must obtain, record and use preservation metadata. This section highlights a small selection of initiatives where metadata is a particular focus.

National Digital Heritage Archive

New Zealand is one of the first countries to extend the legal deposit regime to digital materials. Under the National Library of New Zealand Te Puna Mātauranga o Aotearoa Act 2003, the national library is required to collect, preserve and give access to digital collections. To implement this mandate, they are building a trusted digital repository called the National Digital Heritage Archive (NDHA). The Library expects to select a vendor to provide the software for the NDHA in 2006.[41]

In the meantime, the National Library of New Zealand Te Puna Mātauranga o Aotearoa (NLNZ) has done extensive work related to preservation metadata. The NLNZ schema and data model for preservation metadata mentioned above were released in 2002 and revised in 2003. NLNZ then commissioned a metadata extraction tool from Sytec Resources Ltd, based on the preservation metadata data model. The tool automatically extracts technical preservation metadata from the headers of digital files. It consists of a generic Java application and an "adapter" class for each file format it recognizes. As of late 2005, there were adapters for MS Word 2, MS Word 6, Word Perfect, Open Office, MS Works, MS Excel, MS PowerPoint, TIFF, JPEG, WAV, MP3, HTML, PDF, GIF, and BMP. Other formats can be processed by writing new adapters. Metadata is output in XML and can be transformed by XSL stylesheets into a format that can be loaded into a preservation repository. The tool is designed for use by the wider digital preservation community. It is freely available, along with documentation and a user manual, from the NLNZ website.[42] In 2004 it was a finalist for the Pilgrim Trust's Digital Preservation Award.

MathArc

MathArc (Ensuring Access to Mathematics Over Time) is a collaborative project of the Cornell University Library and Göttingen State and University Library with funding from the National Science Foundation and the Deutsche Forschungsgemeinschaft. The goal of the project is to create a "distributed, interoperable system for the long-term preservation and dissemination of digital serial literature in mathematics and statistics."43 The project architecture is based on the premise, widely accepted by the digital preservation community, that "no single provider can be expected to maintain and ensure access to the archived literature of a single discipline, much less all archived literature. Such an approach would be unworkable technically and financially. Cost effective solutions to digital archiving must share responsibilities of long-term maintenance across numerous stakeholders."[44]

MathArc aims to develop the architecture and workflows to integrate repositories at Göttingen and Cornell into a single preservation and dissemination system conforming to the OAIS Functional Reference Model. A key component of the project is to develop a mechanism for exchanging content and metadata between the partners. MathArc uses the Open Archives

[41] Knight, Steve, October 2005, *In Perpetuity: A Nation's Well-Spring of Knowledge, Library Connect*, v. 3 no 4, http://www.elsevier.com/wps/find/librariansinfo.librarians/LCN030404 [Accessed 22 November 2005, 15:46].

[42] http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction [Accessed 22 November 2005, 16:05].

[43] http://www.library.cornell.edu/dlit/MathArc/web/index.html [Accessed 22 November 2005, 16:06]

[44] http://www.library.cornell.edu/dlit/MathArc/web/resources/projDesc-final.pdf [Accessed 22 November 2005, 16:08]

Initiative Protocol for Metadata Harvesting (OAI-PMH) as the transfer protocol. The unit of transfer is called an "asset," and may be an article, an issue, a volume or even the entire run of a journal. A METS based metadata schema is used to store metadata, structure information and content in a single container.

The METS file includes descriptive metadata in simple Dublin Core, a structural map section describing the logical structure of the asset, and a file description section containing the URIs of all content files included in the asset. Technical metadata is stored in the administrative metadata section of the METS file, and is taken from MIX (Z39.87) for images, a METS extension schema called TextMD for text, and JHOVE schemas for other format types. PREMIS object and event schemas are used for preservation metadata, and are stored in the digital provenance section of the METS schema. Many decisions had to be made to implement PREMIS data elements in this context. PREMIS elements that are not meaningful in this context are given default values. Where METS and PREMIS overlap (for example, for format neutral technical metadata), in some cases the information is stored redundantly while in other cases the information is omitted from the preservation metadata area.[45]

Victorian Electronic Records Strategy (VERS)

The archives and records management professions have developed their own metadata specifications for digital preservation. Because their focus is on the evidentiary value of records, recordkeeping metadata specifications tend to emphasize the authenticity and integrity of electronic records.[46] One of the most interesting active programs is the Victorian Electronic Records Strategy (VERS) of the Public Record Office, Victoria, Australia.[47] VERS uses a three-part approach to the preservation of electronic records, first converting the document to a "long-term format," then encapsulating one or more documents in XML along with their metadata, and finally digitally signing the bundle.[48]

The VERS Metadata Scheme is a superset of the National Archives of Australia's *Recordkeeping Metadata Standard for Commonwealth Agencies* with an emphasis on elements needed for long-term preservation.[49] In archives and records management, a record is a document created or received by an institution, organization, or individual in the course of transacting business or fulfilling a legal obligation. A file is a group of logically associated records, which (if in electronic form) may or may not be physically co-located. Objects encoded in XML are called VERS Encapsulated Objects (VEOs). Record VEOs contain the record content itself and record metadata, and File VEOs contain metadata for computer files, which may include information not found in the individual records.

One of the challenges of this technique is that corrections or enhancements to metadata change the encapsulated object itself. VERS has addressed this with what they call the "onion model." "Changes to a record's metadata may be made by wrapping a new layer of XML around the existing Record VEO. In this way it is possible to store a record's 'history' with the record itself."[50] An excerpt from the VERS specification shows an interesting dependency between preservation strategies and metadata:

[45] Brant, Olaf, Markus Enders, Bill Kehoe, and Marcy Rosenkrantz, *metadata schema for exchanging AIPs, version 1.1,* http://www.library.cornell.edu/dlit/MathArc/web/resources/MathArc_metadataschema031a.doc [Accessed 26 November 2005, 7:33].

[46] Day, Michael, *Preservation Metadata*.

[47] VERS home page, http://www.prov.vic.gov.au/vers/vers/default.htm [Accessed 26 November 2005, 7:37].

[48] Quenault, Howard, *VERS: Building a Digital Record Heritage*, http://www.vala.org.au/vala2004/2004pdfs/13Quena.PDF [Accessed 26 November 2005, 7:39].

[49] *Recordkeeping Metadata Standard for Commonwealth Agencies*, version 1.1, May 1999, http://www.naa.gov.au/recordkeeping/control/rkms/summary.htm [Accessed 26 November 2005, 7:46].

[50] http://www.prov.vic.gov.au/vers/toolkit/resources.htm [Accessed 26 November 2005, 7:48].

> The NAA recommends that records be preserved by migrating them from each system to its replacement. For this reason, the NAA Recordkeeping metadata set maintains a number of types of "history" metadata... which are continually added to over time. The VERS approach, however, is to fix records at (or close to) the time of creation using digital signatures. Although the VERS approach has many advantages over migration, it has one significant disadvantage; metadata that changes or accretes (e.g. use histories) over time is not well supported. Although it is possible to 'layer' metadata to support changing or accreting metadata, this is not efficient for elements that are continually modified.[51]

Web Archiving

The International Internet Preservation Consortium (IIPC) was created in July 2003 by the Internet Archive and eleven national libraries.[52] Led by the Bibliothèque nationale de France, the IIPC is developing tools and interoperability standards to facilitate the archiving of Web content and access to the archived content. The consortium and its partners are working on a standard format for Web archiving and interchange, called Warc, which includes both metadata and content data. They are also working on a standard web archiving metadata set.[53]

The draft IIPC Web Archiving Metadata Set is modeled in a hierarchy of layers, from low-level server interactions, to individual files, to web pages made up of those files, to websites made up of web pages, to crawls archiving multiple websites, ultimately up to the collection. It includes technical metadata similar to that included in PREMIS but also documentation unique to web harvesting, such as information about the crawling tool and selection policy, and the context of the interaction with the web server. A first draft of the metadata specification is expected to be available for public comment in November 2005.[54]

PRONOM

Most specialists agree that complex information that is common to all objects of a particular type or format might best be supplied by automated look-up in a central registry. For example, all PDF 1.6 files are binary, big-endian, and documented by the same Adobe reference manual. If this information were stored in a central place, there would be no need for individual repositories to record these details so long as they recorded the file format as PDF 1.6 and knew where to find the additional information. This is the theory behind a number of central registries in various stages of development.

The most mature preservation registry is PRONOM, an online service of the UK National Archives.[55] PRONOM stores information about file formats, software products, software vendors, and product support providers. At this time PRONOM is designed for human-interactive use, and can not be queried by program. For a given file format, PRONOM will list detailed technical metadata and a list of software applications that use or render the format. A search of "PDF" for example, returns a list of PDF versions from 1.0 to 1.6. Selecting a version results in a display of summary specifications, links to authoritative documentation, "signatures" (or characteristics by which the format can be identified) and other information. PRONOM can be searched a number of ways: by file format, by software

[51] *Standard for the Management of Electronic Records PROS 99/007, version 1, Specification 2, VERS Metadata Scheme*, http://www.prov.vic.gov.au/vers/standard/ver1/99-7-2s2.htm [Accessed 26 November 2005, 7:53].

[52] International Internet Preservation Homepage, http://netpreserve.org [Accessed 26 November 2005, 7:54].

[53] Lupovici, Catherine, 2005, *Web archives long term access and interoperability: the International Internet Preservation consortium activity*, http://www.ifla.org/IV/ifla71/papers/194e-Lupovici.pdf [Accessed 26 November 2005, 7:56].

[54] Masanès, Julien, 2005, IIPC Web Archiving Metadata Set http://www.iwaw.net/05/masanes2.pdf [Accessed 26 November 2005, 7:58].

[55] PRONOM home page, http://www.nationalarchives.gov.uk/pronom/ [Accessed 26 November 2005, 8:00].

product, by vendor, and by PRONOM identifier. An interesting search called "lifecycle" will show all software products supported as of a certain date, or released before or after a given date.

Most information to date has been supplied by digital preservation staff at the National Archives, working with major software developers. However, developers of software products and file format specifications are encouraged to submit information directly to PRONOM. The National Archives intends to use the PRONOM site to provide a suite of tools as well as registry information, and recently released DROID, a free tool for identifying file formats.

Next steps

Any organization thinking about implementing a long-term preservation repository must think about preservation metadata. This includes not only information about the materials that will be stored in the repository, but also event or tracking information for pre-ingest and repository actions, and information about ownership, rights and permissions. As explained by Lavoie and Gartner,

"The scope and depth of the preservation metadata required for a given digital preservation activity will vary according to numerous factors, such as the 'intensity' of preservation, the length of archival retention, or even the knowledge base of the intended user community."[56]

This means there is no one-size-fits-all solution, and every repository must make and understand its own decisions.

The archive may plan to use a locally-developed preservation repository application, an application developed in response to an RFP or specifications drafted by the organization, or an open-source or vendor-provided application. In the first two cases, the organization should have considerable metadata expertise on staff and/or available as consultants. Many elements of preservation metadata will be used by the repository application itself in order to perform preservation functions, so specifications for metadata and for software must go hand in hand. The data curators will want to review the PREMIS Data Dictionary and other published metadata specifications and adopt those elements applicable to the type of archive and the type of material. It should be possible for the repository to export PREMIS conformant metadata to facilitate interoperability with other repositories.

If the organization is selecting a third-party repository application, the amount and nature of the preservation metadata to be recorded has probably been determined by the application. In this case the metadata analysis should be part of the process of selecting the application. The archiving organization should feel confidant that the metadata maintained by the application is sufficient to carry out the expected preservation functionality.

The organization should also consider the way that metadata is stored. There is a growing consensus that archived objects should be self-documenting, in the sense that they should be stored with all the information needed to identify, understand, and use them. In practice, this means that both descriptive and preservation metadata should be stored along with the content the metadata pertains to, regardless of whether it is also stored in more accessible form for use by the repository application. The DAITSS system under development by the Florida Center for Library Automation, for example, stores complete preservation metadata in XML along with each

[56] Lavoie, Brian and Richard Gartner, *Preservation Metadata*, p. 2.

archived object as well as in a relational database for fast access.[57]

In sum, whether the archiving organization is developing its own application or evaluating existing systems, it should consider the following questions:

- To what extent is metadata obtained in an automated way as opposed to requiring user submission or input?
- Can the repository demonstrate the fixity of archived files?
- Will the repository be able to document the authenticity of archived materials, starting at least at the point of ingest, by maintaining a detailed record of events in the lifecycle of the object?
- Can the repository record actions and technical characteristics for both files and conceptual objects (meaningful aggregations of files)?
- Is it important that the repository maintain metadata for controlling materials at a higher level (collections) or lower level (bit streams)?
- What preservation strategies (migration, normalization, emulation, cannonicalization, etc.) will the system implement; how will it use metadata in this process?
- Does the system save metadata in archival storage along with content objects, as well as keeping a working copy to support repository operations?
- Will the repository be able to export standards-conformant metadata according to published XML schema?

Future developments

Despite the impressive amount of effort that has been devoted to preservation metadata over the last decade, a great deal remains to be done. Developments to watch fall into two categories: the standardization and refinement of preservation metadata element specifications, and the incorporation of more preservation metadata into more repository applications.

The set of core elements in the PREMIS Data Dictionary has been widely accepted, at least in principle, but it has not yet proven itself through experience in operational repositories. We can expect a number of case studies to report on both implementation and use in carrying out preservation strategies. Projects such as MathArc will provide information on the utility of the PREMIS set for inter-repository exchange.

Standardized metadata element sets are still needed for technical metadata for nearly all formats except still images. Applications such as JHOVE and the NLNZ metadata extraction tool which extract technical metadata from file headers may expedite the development of standards in this area.

There is little understanding of the intellectual property rights needed for long-term digital preservation, including which permissions are needed in order to carry out preservation strategies such as migration and emulation. As this area is clarified, the definition of metadata schema for recording rights and permissions will surely follow. It is not clear at this time whether rights expression languages proposed for other uses have utility in the preservation setting.

In the future there may also be some attempts to integrate the various aspects of preservation metadata into more comprehensive standards for particular types of materials. For example, the *Digital Images Archiving Study* released by the Arts and Humanities Data Service proposes the integration of PREMIS, Z39.87 (Technical

[57] Caplan, Priscilla, 2005, *Building a Dark Archive in the Sunshine State: A Case Study*, http://www.fcla.edu/digitalArchive/pdfs/IS_Tpaper.pdf [Accessed 26 November 2005, 9:10].

Metadata for Digital Still Images) and METS into a single framework for use by image archives.[58]

The national libraries and national archives that are implementing preservation repository applications tend to record detailed metadata and make extensive use of it. Repository applications that are more generally available as vendor or open source software have to date made less sophisticated use of preservation metadata. It is likely that applications intended for use as institutional repository systems will evolve to maintain more preservation metadata, at the same time as applications specifically devoted to digital preservation will be come available.

Conclusions

Preservation metadata is the information necessary to support the process of digital preservation over the long-term. As such it is not an end in itself, but a means to accomplishing the ends of sustainability and long-term usability of digital collections. Metadata must document the technical characteristics and significant properties of the objects being preserved sufficiently to support the repository in carrying out its chosen preservation strategies. It must document ownership and intellectual property sufficiently to allow the repository to undertake preservation actions. It must document the origin, fixity, and provenance of objects sufficiently to support claims of authenticity.

Preservation metadata is, on the whole, not simple to understand, obtain, or implement. Because the amount of desirable metadata can be extensive and its accuracy is important, the metadata should be obtained automatically whenever possible. There are a growing number of tools available for extracting technical metadata from digital files in various formats. Some technical and environment information will also be available in centrally maintained registries. However, much of the information that documents provenance, the structure of objects, relationships between objects, preservation actions, and intellectual property rights may be more difficult to obtain.

Many specifications for preservation metadata have been published and significant progress has been made towards standardizing a core set of preservation metadata elements. However, because digital preservation repositories are a relatively new phenomenon, the success of preservation metadata in supporting long-term preservation is largely untried. The metadata recorded today is our best guess of what will be useful tomorrow. As more experience is gained with various preservation strategies and different preservation repository systems, we can expect our understanding of preservation metadata to grow increasingly more sophisticated in the future.

[58] Arts and Humanities Data Service, 2006, *Digital Images Archiving Study* Draft Report, http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf [Accessed 22 April 9:31].

Terminology

Administrative metadata
Metadata primarily intended to facilitate the management of resources.

Authenticity
The property that an object is what it purports to be; that the source and content of the object are as represented.

Fixity
The property of being unchanged between two points in time.

Digital provenance
The origin and chain of custody of a digital object, and the history of events affecting it.

Preservation metadata
Metadata that supports and documents the process of digital preservation.

Preservation repository
See Repository.

Repository
A facility for storing and maintaining digital resources. A preservation repository is one whose mission is to maintain resources over the long-term by applying one or more preservation strategies such as normalization, migration or emulation.

Structural metadata
Metadata that describes the internal organization of a digital resource.

Technical metadata
Metadata primarily intended to document the creation and characteristics of digital files. Some elements of technical metadata, for example format and size, pertain to all files. Other elements of technical metadata are format-specific. For example, character set pertains only to text files, while frame rate pertains only to video files.

References

Consultative Committee for Space Data Systems, January 2002, *Reference Model for an Open Archival Information System (OAIS)*, http://www.ccsds.org/documents/650x0b1.pdf [Accessed 22 November 2005, 12:59].

Day, Michael, *Preservation metadata. Prepublication draft of chapter published in: G. E. Gorman and Daniel G. Dorner (eds.), Metadata applications and management*, International Yearbook of *Library and Information Management, 2003-2004, London: Facet Publishing, 2004, pp. 253-273,* http://www.ukoln.ac.uk/metadata/publications/iylim-2003/ [Accessed 22 November 2005, 14:13].

Data Dictionary - Technical Metadata for Digital Still Images (DRAFT), http://www.niso.org/standards/resources/Z39-87-200x-forballot.pdf?CFID=6860130&CFTOKEN=81464797 [Accessed 22 November 2005, 14:50]

Knight, Steve, October 2005, *In Perpetuity: A Nation's Well-Spring of Knowledge, Library Connect*, v. 3 no 4, http://www.elsevier.com/wps/find/librariansinfo.librarians/LCN030404 [Accessed 22 November 2005, 15:46].

Lavoie, Brian and Richard Gartner, September 2005, *Technology Watch Report: Preservation Metadata*, http://www.dpconline.org/docs/reports/dpctw05-01.pdf [Accessed: 22 November 2005, 12:23].

Lord, Philip, and Alison Macdonald, 2003, *Data Curation for e-Science in the UK: An Audit to Establish Requirements for Future Curation and Provision*, http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Lupovici, Catherine, 2005, *Web archives long term access and interoperability: the International Internet Preservation consortium activity*, http://www.ifla.org/IV/ifla71/papers/194e-Lupovici.pdf [Accessed 26 November 2005, 7:56].

Lupovici, Catherine and Julien Masanès, July 2000, *Metadata for Long-term Preservation*, http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm [Accessed 22 November 2005, 14:05].

Masanès, Julien, 2005, IIPC Web Archiving Metadata Set http://www.iwaw.net/05/masanes2.pdf [Accessed 26 November 2005, 7:58].

Masanès, Julien, May 2003, *Technical Information Needed for Long-term Preservation of Digital Documents*, in International Preservation News v. 29, http://www.ifla.org/VI/4/news/ipnn29.pdf [Accessed 26 November 2005, 11:32].

Metadata for Digital Preservation: The CEDARS Project Outline Specification Draft for Public Consultation, March 2000, http://www.leeds.ac.uk/cedars/cedars.pdf [Accessed 22 November 2005, 13:59].

National Library of Australia, 15 October 1999, *Preservation Metadata for Digital Collections*, http://www.nla.gov.au/preserve/pmeta.html [Accessed 22 November 2005, 13:56].

National Library of New Zealand, July 2003, *Metadata Standards Framework – Metadata Implementation Schema*, http://www.natlib.govt.nz/files/nlnz_data_model.pdf [Accessed 22 November 2005, 14:15].

National Science Board, October 2005, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf [Accessed 22 November 2005, 15:20].

OCLC/RLG Working Group on Preservation Metadata, 31 January 2000, *Preservation Metadata for Digital Objects: A Review of the State of the Art*, http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf [Accessed 22 November 2005, 14:08].

OCLC/RLG Working Group on Preservation Metadata, June 2002, *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*, http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf [Accessed 22 November 2005, 14:10].

Public Records Office Victoria, Management of Electronic Records (PROS 99/007), http://www.prov.vic.gov.au/vers/standard/ [Accessed 26 November 2005, 10:27].

PREMIS Working Group, May 2005, *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*, http://www.oclc.org/research/projects/pmwg/premis-final.pdf [Accessed 22 November 14:24].

Quenault, Howard, *VERS: Building a Digital Record Heritage*, http://www.vala.org.au/vala2004/2004pdfs/13Quena.PDF [Accessed 26 November 2005, 7:39].

Recordkeeping Metadata Standard for Commonwealth Agencies, version 1.1, May 1999, http://www.naa.gov.au/recordkeeping/control/rkms/summary.htm [Accessed 26 November 2005, 7:46].

Searl, Sam and Dave Thompson, *Preservation Metadata: Pragmatic First Steps at the National Library of New Zealand*, http://www.dlib.org/dlib/april03/thompson/04thompson.html [Accessed 22 November 2005, 14:18].

Waters, Donald and John Garrett, 1996, *Preserving Digital Information: Final Report of the Task Force on Archiving of Digital Information*, ftp://ftp.rlg.org/pub/archtf/final-report.pdf [Accessed: 22 November 2005, 12:51].