Technology Watch Report

Preservation Metadata

Brian Lavoie
Office of Research
OCLC Online Computer Library Center, Inc.
lavoie@oclc.org

Richard Gartner
Oxford University Library Services
richard.gartner@sers.ox.ac.uk

DPC Technology Watch Series Report 05-01
September 2005

Keywords: metadata, digital preservation, schema, OAIS, metadata framework, PREMIS, data dictionary, METS, XML schema, archival information package

Executive Summary

Preservation metadata is information that supports and documents the long-term preservation of digital materials. It addresses an archived digital object's *provenance*, documenting the custodial history of the object; *authenticity*, validating that the digital object is in fact what it purports to be, and has not been altered in an undocumented way; *preservation activity*, documenting the actions taken to preserve the digital object, and any consequences of these actions that impact its look, feel, or functionality; *technical environment*, describing the technical requirements, such as hardware and software, needed to render and use the digital object; and *rights management*, recording any binding intellectual property rights that may limit the repository's ability to preserve and disseminate the digital object over time. Preservation metadata addresses all of these issues and more. In short, preservation metadata helps make an archived digital object self-documenting over time, even as the intellectual, economic, legal, and technical environments surrounding the object are in a constant state of change. The principal challenge in developing a preservation metadata schema is to anticipate what information will actually be needed to support a particular digital preservation activity, and by extension, to meet a particular set of preservation goals.

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.

The OAIS (Open Archival Information System) reference model provides a high-level overview of the types of information needed to support digital preservation, including representation information, preservation description information (which can be broken down into reference, context, provenance, and fixity information), packaging information, and descriptive information. These information types can be interpreted as the general categories of metadata needed to support the long-term preservation and use of digital materials, and have served as the starting point for a number of preservation metadata initiatives.

Over the past few years, a number of institutions and projects have released preservation metadata element sets, reflecting a wide range of assumptions, purposes, and approaches. In 2002, the OCLC/RLG Preservation Metadata Framework Working Group consolidated existing expertise in the form of a preservation metadata framework. Using the broad categories of information specified in OAIS as a starting point, the Framework enumerates the types of information falling within the scope of preservation metadata. The working group then expanded each category of information, providing additional structure to articulate the OAIS information requirements in progressively greater detail and ending with a set of "prototype" preservation metadata elements.

Release of the Framework prompted interest in moving it toward a more implementable status. In response to this, OCLC and RLG sponsored a second working group: PREMIS (PREservation Metadata: Implementation Strategies). Composed of more than thirty international experts in preservation metadata, PREMIS sought to: 1) define a core set of implementable, broadly applicable preservation metadata elements, supported by a data dictionary; and 2) identify and evaluate alternative strategies for encoding, storing, managing, and exchanging preservation metadata in digital archiving systems. In September 2004, PREMIS released a survey report describing current practice and emerging trends associated with the management and use of preservation metadata to support repository functions and policies. PREMIS followed up the survey report with the 237-page *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*, released in May 2005. The PREMIS Data Dictionary is a comprehensive, practical resource for implementing preservation metadata in digital archiving systems. It defines implementable, core preservation metadata, along with guidelines and recommendations for management and use. A maintenance activity has been set up to manage the Data Dictionary and coordinate future revisions.

Digital objects accumulate a great deal of metadata over time. METS (Metadata Encoding and Transmission Standard) is an XML schema designed specifically as an overall framework within which all the metadata associated with a digital object can be stored. A METS file comprises four major constituent sections: a file inventory for all the files associated with the digital object; a section for administrative metadata; a section for descriptive metadata; and a structural map for the object. METS allows two approaches to the storage of the metadata and data associated with a digital object: both may be either stored internally within the METS file, or held externally and referenced from within METS. The content of each section is not prescribed by METS itself: any XML data or metadata may be used; however, METS does recommend a number of schemas. The flexibility of METS implies that its practical implementation can be very flexible as well: any system capable of handling XML documents can be used to create, store and deliver METS-based metadata. METS Profiles can be used to document a particular METS implementation within a project.

The resources required for a METS-based system are no more than one requires for handling any other form of XML object. METS is at its strongest when dealing with a wide variety of materials which need to be handled flexibly but in an organized and coherent manner. Since XML is non-proprietary, a METS file is not tied to any given software package, which mitigates the threat of technical obsolescence. Because METS was designed to act as an OAIS Archival Information Package, no conceptual leap is required to fit METS into the OAIS landscape.

There are a number of areas future preservation metadata work could address. Automated preservation metadata tools are needed. Ideally, these tools should support formal preservation metadata schema, and be surfaced in a variety of digital asset management environments. Collaborative metadata management strategies, such as the sharing and re-use of preservation metadata through registries, and diffusion of metadata capture responsibilities throughout the digital information lifecycle, can offer efficient, economical ways of acquiring and maintaining certain forms of preservation metadata. Finally, there is a need to explore the implications of exchanging preservation metadata across a network of heterogeneous digital archiving systems.

DPC Technology Watch Report on Preservation Metadata[1]

Metadata and preservation metadata

It is hard to discuss information management topics today without encountering the term *metadata*. The canonical definition of metadata – "data about data" – is not particularly helpful in understanding what metadata is and how it is used, but fortunately, more informative definitions are available. Metadata definitions resemble the old adage about standards – the nice thing is there are so many to choose from – but a good one is provided by the National Information Standards Organization[2], who define metadata as "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource".

'Metadata definitions resemble the old adage about standards – the nice thing is there are so many to choose from...'

Most metadata primers slice up metadata into three distinct categories. *Descriptive metadata* is information that identifies, supports the discovery of, and documents relationships between, information resources. *Structural metadata* is information that documents how the component pieces of complex information resources fit together, such as the chapters in a book, or the text, image, and sound files that collectively make up a multi-media Web page. *Administrative metadata* is information that supports one or more processes concerned with the management of information resources. Caplan[3] provides a detailed discussion of descriptive, structural, and administrative metadata.

Segmenting metadata into descriptive, structural, and administrative categories is helpful for thinking about the range of information potentially encapsulated in a metadata schema[4], but at the same time, it can be misleading to the extent that it suggests clear boundaries exist between the three categories. In fact, the distinction between descriptive, structural, and administrative metadata can be quite blurred. Even more problematic are attempts to assign a metadata schema to a single category, since most tend to span multiple categories. Nowhere is this problem more evident than in endeavoring to set *preservation metadata* in the broader context of the general metadata landscape.

Preservation metadata is information that supports and documents the long-term preservation of digital materials. Many commentators assign preservation metadata to the category of administrative metadata, since preservation is an information management process – but in fact, this categorization is not correct. A preservation metadata schema will include descriptive, structural, and administrative metadata elements. In light of this,

[1] The authors wish to thank two anonymous reviewers for their comments and suggestions.

[2] NISO (2004) *Understanding Metadata*, p.1. Available at: http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

[3] Caplan, P. (2003) *Metadata Fundamentals for All Librarians* (Chicago: ALA)

[4] For the purposes of this paper, a *metadata schema* may be understood as a set of metadata elements along with guidelines or instructions for their use.

we must look for criteria for distinguishing preservation metadata from other forms of metadata at a level somewhere above the descriptive/structural/administrative distinction.

A metadata schema is intended to serve some purpose – in a sense, it supports some “verb” that the schema is intended to help accomplish. In the case of preservation metadata, the verb is “preserve”, so when we draw a distinction between metadata and preservation metadata, it is this verb which helps us draw a boundary around what is in and what is out of scope. *Preservation metadata is descriptive, structural, and administrative metadata that supports the long-term preservation of digital materials.*

Requirements and importance

It is difficult to draw a clear boundary around what types of information fall within the scope of preservation metadata. It is probably too much to say that preservation metadata is any metadata used in a digital preservation repository setting, yet it is certainly more than the technical information needed to maintain and render digital formats across changing technology cycles. There has, however, been much discussion of what types of information are required to support the digital preservation process, and from this, consensus seems to have settled around five major areas relevant to preservation metadata:

> **‘It is probably too much to say that preservation metadata is any metadata used in a digital preservation repository setting, yet it is certainly more than the technical information needed to maintain and render digital formats across changing technology cycles.’**

Provenance: Preservation metadata should record information bearing on the custodial history of the digital object, potentially stretching back to the time of the object’s creation, and moving forward through successive changes in physical custody and/or ownership.

Authenticity: Preservation metadata should include information sufficient to validate that the archived digital object is in fact what it purports to be, and has not been altered, either intentionally or unintentionally, in an undocumented way.

Preservation activity: Preservation metadata should document the actions taken over time to preserve the digital object, and record any consequences of these actions that impact the look, feel, or functionality of the object.[5]

Technical environment: Preservation metadata should describe the technical requirements, such as hardware, operating system, and software applications, needed to render and use the digital object in the state in which it is currently stored in the repository.

Rights management: Preservation metadata should record any binding intellectual property rights that limit the repository’s powers to take action to preserve the digital object, and to disseminate the object to current and future users.

[5] Documentation of the nature and impact of digital preservation activities could be construed as a form of provenance information; however, it is of sufficient importance to merit separate emphasis.

It is readily seen that the information types enumerated above constitute a deep, comprehensive description of the custodial, technical, and even legal aspects of archived digital objects. This is a great deal of information to collect and maintain, which naturally invites the question of why it is necessary to do so – in other words, why is preservation metadata important?

First, preservation metadata is important because *digital objects are technology-dependent*. Unlike print books or oil paintings, the contents of digital objects cannot be accessed "directly" by users; instead, a complex technological environment, consisting of software, hardware, and in some cases network technology, sits between the user and the object's contents. Rendering and using digital objects requires the availability of this environment, or at least some technically equivalent substitute. For this reason, it is not enough to simply preserve a digital object: the means to render and use it must be preserved as well. This need is amplified in light of the constant pace of technological change, which inevitably makes today's technologies obsolete. Consequently, it is especially important to carefully document the technological environment of an archived digital object to ensure it remains usable for current and future generations.

'…it is not enough to simply preserve a digital object: the means to render and use it must be preserved as well.'

A second reason preservation metadata is important is that *digital objects are mutable*. They can be easily altered, either by accident or design, with potentially significant consequences for an object's look, feel, and functionality. Beyond this, the relatively short lifespan of many forms of digital storage media raise the specter of "bit rot" – the gradual degradation of stored bits leading to partial or even complete information loss. Even the act of preservation itself can alter the form or function of a digital object – for example, when an object is migrated from one format to another in order to keep pace with changing technologies. For these and other reasons, it is especially important for an archived digital object to be accompanied by metadata documenting its provenance and authenticity – in particular, its salient characteristics at the time of creation, how those characteristics have been altered over time, by whom, and for what purpose. This becomes especially important in domains such as electronic record-keeping, where the evidentiary value of the content must be preserved and validated.

Finally, preservation metadata is important because *digital objects are bound by intellectual property rights*.[6] The relatively brief "shelf life" of digital storage media, along with the rapid obsolescence of contemporary technology, often produces a very short "window of inactivity" during which preservation actions can be safely deferred.

[6] This is not to say that non-digital objects are not bound by IPR, but there is an important distinction between the two formats. For non-digital objects – e.g., print materials – preservation actions can often be deferred for a considerable period of time; the process of degradation is slow enough that by the time preservation actions become imperative, the material has either passed into the public domain, or its owners have, for one reason or another, relinquished their rights attached to the object – perhaps because the object has ceased to hold a private economic value. In these circumstances, public agencies are often free to intervene and take whatever actions are necessary to preserve the object over the long-term.

Moreover, digital preservation actions are, for the most part, pre-emptive in nature, seeking to avert damage rather than to repair it. Once a digital file is corrupted, or the means to access it lost, its contents may be lost forever. In light of these considerations, digital preservation must often take place early in the information life cycle – and while the material is still under copyright. So rather than operating with a free hand, preservation repositories often must work within limitations imposed by currently binding property rights that define acceptable preservation and access policies. The impact of intellectual property rights on digital preservation can vary across contexts, and be manifested in complex ways – for example, even if the archived content is in the public domain, rights may still be attached to the software needed to render it. For these reasons, it is especially important to document the intellectual property rights associated with an archived digital object, in order that long-term preservation actions can be coordinated with any rights restrictions binding on the object.

There are many other reasons why preservation metadata is an important – indeed an essential – component of most digital preservation strategies. A useful way of summing them all up might be as follows: preservation metadata is important because it enables a digital object to be *self-documenting* over time, and therefore positioned for long-term preservation and access, even as ownership, custody, technology, legal restrictions, and even user communities are relentlessly changing.

'…preservation metadata is important because it enables a digital object to be *self-documenting* over time, and therefore positioned for long-term preservation and access…'

Developing a preservation metadata schema[7]

The principal challenge in developing a preservation metadata schema is to anticipate what information will actually be needed in order to support a particular digital preservation activity, and by extension, to meet a particular set of preservation goals. 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 (i.e., whether an archived object's intellectual content can be migrated to new formats to keep pace with changing technology, or whether the object must be maintained in a form that preserves its original look, feel, and functionality); the length of archival retention (e.g., finite, as in the case of legal obligations to maintain institutional records for a prescribed period; or "in perpetuity", as in the case of a digital resource that is part of the permanent historical record); or even the knowledge base of the intended user community. Perhaps more than any other form of metadata, preservation metadata requires planners to "get it right" the first time.

'Perhaps more than any other form of metadata, preservation metadata requires planners to "get it right" the first time.'

Once a preservation metadata schema has been developed and implemented, it is difficult to judge its effectiveness *a priori*.

[7] Parts of this section are adapted from Lavoie, B. (2004) "Preservation Metadata: Challenge, Collaboration, and Consensus" *Microform & Imaging Review* Vol. 33, No. 3.

Metadata intended to aid resource discovery can be readily tested, and if necessary, refined, in order to improve the relevance and accuracy of search results. In contrast, the suitability of a particular set of preservation metadata elements may not be determined until long after their implementation, at which time a digital repository might discover that the metadata collected far exceeds what was actually necessary, or conversely (and more serious), was insufficient to support the long-term requirements of the digital archiving system. [8]

Many factors need to be taken into account when developing a preservation metadata schema, but three are of particular importance. These factors may seem obvious, but are nevertheless worth stating explicitly.

A preservation metadata schema should aim to be:

- *Comprehensive:* Even if the scope and depth of the schema exceeds current needs, it is easier to employ a limited set of elements from the schema now, and preserve the option to adopt other portions of the schema later as need arises, than to fully employ a limited schema now, and be forced to extend the schema in an ad hoc, "piece-meal" fashion, should it be determined over time that additional information is needed.

- *Oriented toward implementation:* Metadata is expensive to create and maintain; a good preservation metadata schema should, therefore, be designed with the practicalities of implementation in mind. For example, the schema should, where possible, provide controlled vocabularies or codes for populating elements, rather than relying on "free text". In addition, the schema should be adaptable to automated workflows for metadata collection and management; see below for more on this point.

- *Interoperable:* The "entourage" of metadata that accompanies a digital object over time will likely be accumulated from, and used by, a variety of stakeholders beyond the repository itself. Given this, a preservation metadata schema should be designed to promote interoperability across these stakeholders, in the sense of facilitating transactions involving an archived digital object and/or its associated metadata: e.g., initial submission to the repository, dissemination to a user, or transfer to another repository.

OAIS to PREMIS, or, preservation metadata from theory to practice

Although it is still a fairly new topic, preservation metadata has moved quite rapidly from theory to practice. In part, this mirrors overarching conditions in the digital preservation area itself, where efforts to carefully develop solid foundations for digital preservation

[8] More generally, one could argue that metadata schema are products of the time and conditions under which they were produced, and therefore are subjective to some degree. For a discussion of this point, see Bowker, G. C. (2000) "Biodiversity Datadiversity" *Social Studies of Science* Vol. 30 No.5, p.643-683. The authors thank an anonymous reviewer for this point.

techniques and practices are paralleled by an immediate need to implement capacity to secure the long-term retention of digital materials currently perceived to be at risk. In this sense, the movement from theory to practice in preservation metadata cannot be traced as a straight line, but rather as a series of overlapping initiatives straddling research and development, with a substantial dose of cross-fertilization at the boundary. For expositional purposes, however, it is useful to establish two endpoints for the development of preservation metadata – the OAIS Information Model at one end, and the PREMIS Working Group at the other – with a number of important initiatives taking place in between.

OAIS

The OAIS (Open Archival Information System) reference model[9] is a conceptual framework describing the environment, functional components, and information objects associated with a system responsible for the long-term preservation of digital materials.[10] OAIS was approved as ISO standard 14721 in 2002, but even before then, it had enjoyed widespread adoption in the digital preservation community. It is common for digital preservation repositories to bill themselves as "OAIS-compliant", although there is no definitive articulation of what such compliance requires.[11]

OAIS has exerted a great deal of influence in the development of the art and science of digital preservation, with preservation metadata one of the areas where this impact has been especially evident. In particular, the *OAIS information model* has served as the foundation for, or at least informed, the development of most preservation metadata initiatives that have emerged in recent years. Indeed, one could argue that the salient characteristic shared by these initiatives, and therefore the starting point for consensus-building in the area of preservation metadata, is the fact that each can be traced, in some form or another, back to the common antecedent of the OAIS information model.

'…there is a fundamental link between preserved digital content and metadata, or put another way, metadata plays an essential role in preserving digital content and supporting its use over the long-term.'

The OAIS information model is a conceptualization of the information objects taken into, stored, and disseminated by a digital preservation repository. The core concept underlying this model is that of an *information package* – a combination of some piece of content that is the focus of preservation, along with its associated metadata. OAIS defines three varieties of information package: a *submission information package*, or SIP, which is the content and associated metadata "ingested" into the repository at the time of deposit; the *archival information package*, or AIP, which is the content and associated

[9] http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf

[10] Strictly speaking, OAIS is intended to represent the conceptual foundations of *any* system tasked with long-term preservation – therefore, the objects that are the focus of preservation could be anything from print books to geological samples. But it is in the context of *digital* preservation that OAIS has received the most attention and take-up.

[11] For a description of the OAIS model, including its history and development, see Lavoie, B. (2004) *The Open Archival Information System Reference Model: Introductory Guide* DPC Technology Watch Report. Available at: http://www.dpconline.org/docs/lavoie_OAIS.pdf

metadata actually stored and managed by the repository over the long-term; and lastly, the *dissemination information package*, or DIP, which is the content and associated metadata provided by the repository in response to access requests by users.

Differentiation across these three package types can include the form of the content, the form of the metadata, or what is more likely, both. But the key point is that regardless of package type, there is a fundamental link between preserved digital content and metadata, or put another way, metadata plays an essential role in preserving digital content and supporting its use over the long-term.

The OAIS information model implicitly establishes the link between metadata and digital preservation – i.e., *preservation metadata*. In addition, it provides a high-level overview of the types of information that fall within the scope of preservation metadata, including:[12]

- Representation Information: information necessary to render and understand the bit sequences constituting the archived digital object.
- Preservation Description Information: information that supports and documents the preservation of the archived object, including:
  Reference information: uniquely identifies the archived object;
  Context information: describes the archived object's relationship(s) to other archived objects;
  Provenance information: documents the history of the archived object;
  Fixity information: validates the authenticity or integrity of the archived object.
- Packaging Information: information that binds all components of an information package into a single logical unit.
- Descriptive Information: information that supports the discovery and retrieval of the archived object by the repository's users.

These information types can be collectively interpreted as the most general description of the metadata needed to support the long-term preservation and use of digital materials. They would serve as the starting point for most subsequent efforts to develop formal preservation metadata schema.

Preservation metadata element sets

As the need to develop operational digital preservation capacity began to surface, a number of institutions undertook to develop preservation metadata element sets to support current or planned efforts to preserve digital materials. There is no space in this paper to attempt an exhaustive list of these element sets, but it is useful to briefly mention several, in order to provide examples of how institutions have implemented preservation metadata requirements in practice, and to convey a sense of the "state-of-the-art" prevailing at the time preservation metadata consensus-building efforts (to be discussed in the next two sections) began to coalesce.

[12] For a more detailed description of these information types, see Lavoie (2004).

Early efforts to develop preservation metadata element sets were undertaken by the National Library of Australia (NLA), the CEDARS (CURL Exemplars in Digital Archives) project, and the NEDLIB (Networked European Deposit Library) project.[13] The NLA element set[14] was designed to support the preservation of both digitized and born-digital objects. It accommodates three levels of descriptive granularity – collection, object, and sub-object (file) – and is implementation-neutral, in the sense that no assumptions are made about the specific preservation strategy adopted by the repository. The CEDARS element set[15] was developed for use with a pilot digital archive, and is applicable to a variety of digital formats. In contrast to the NLA set, these elements are applicable at any level of description. Finally, the NEDLIB element set[16] defines a "core" set of essential preservation metadata, with an emphasis on overcoming the problem of technological obsolescence. Elements are defined at a high level to maximize applicability across object formats and types.

'In considering these and other preservation metadata element sets, one can sum them up by observing that the earlier efforts …largely were speculative in nature, seeking to anticipate the metadata needs of programmatic digital preservation initiatives that would emerge in the future.'

Examples of more recent efforts to develop preservation metadata element sets include those produced by OCLC, the National Library of New Zealand (NLNZ), and the University of Edinburgh (UE). The OCLC element set[17] was developed for use in conjunction with its Digital Archive service; its design benefited from input provided by users of the service. The NLNZ element set[18] supports the Library's ongoing efforts to develop internal digital preservation capacity. It is a starting point for implementing systems responsible for collecting and managing preservation metadata. The UE element set[19] is part of a wider effort to develop a digital preservation strategy for the University and its stakeholders, and is based largely on the CEDARS element set mentioned above. In considering these and other preservation metadata element sets, one can sum them up by observing that the earlier efforts – NLA, CEDARS, NEDLIB, and others – largely were speculative in nature, seeking to anticipate the metadata needs of programmatic digital preservation initiatives that would emerge in the future. On the other hand, development of the more recent element sets, such as OCLC, NLNZ, and UE, were more closely aligned with planning and implementation of "production" digital archiving systems – and of course, benefited considerably from the foundations laid by the earlier sets.

[13] The following descriptions of the NLA, CEDARS, and NEDLIB preservation metadata element sets are adapted from OCLC/RLG (2001) *Preservation Metadata for Digital Objects: A Review of the State of the Art*, p.17-18. Available at: http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf

[14] http://www.nla.gov.au/preserve/pmeta.html

[15] http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html

[16] http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm

[17] http://www.oclc.org/digitalarchive/about/works/metadata/

[18] http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf

[19] http://www.lib.ed.ac.uk/sites/digpres/metadataschema.shtml

Preservation metadata framework working group[20]

The ubiquity of the digital preservation problem speaks to the value of collaboration and consensus-building for resolving the challenges and uncertainties of managing digital materials over the long-term. Digital preservation is an issue that impacts a variety of stakeholders, distributed throughout the academic, commercial, government, and cultural heritage communities, and each confronted with a similar need to develop effective strategies for securing the long-term retention of digital materials. In 2000, OCLC and RLG jointly sponsored the creation of an international working group tasked with defining the role of metadata in the digital preservation process. The Preservation Metadata Framework Working Group[21] drew together the expertise of individuals from a variety of institutional backgrounds.

'The ubiquity of the digital preservation problem speaks to the value of collaboration and consensus-building for resolving the challenges and uncertainties of managing digital materials over the long-term.'

At the time the working group was organized, there was little or no consensus on even the most fundamental questions surrounding preservation metadata, including what types information constituted preservation metadata, and how it could be used to support the digital preservation process. As discussed in the previous section, several institutions had developed element sets for internal use, but these reflected a wide range of assumptions, purposes, and approaches. In light of this, the working group produced a white paper[22] summarizing the "state of the art" in preservation metadata. The white paper provided a definition of preservation metadata, described its role in the digital preservation process, and reviewed a number of existing preservation metadata initiatives, with an emphasis on identifying points of convergence and divergence among them.

The white paper provided a foundation for the working group's next task, which was to develop a comprehensive, broadly applicable *preservation metadata framework* enumerating the types of information falling within the scope of preservation metadata. Given its extensive take-up in the digital preservation community, the working group chose OAIS as the starting point for the framework. The broad categories of information specified in the OAIS information model served as a top-level description of the types of information comprising preservation metadata. The working group then expanded each category of information, providing additional structure to articulate the OAIS information requirements in progressively greater detail and ending with a set of "prototype" preservation metadata elements.

Published in 2002, the preservation metadata framework[23] was the first international consensus-driven statement on the scope of preservation metadata. It consolidated existing expertise to create a solid foundation upon which future preservation metadata

[20] Parts of this section are adapted from Lavoie, B. (2004) "Preservation Metadata: Challenge, Collaboration, and Consensus" *Microform & Imaging Review* Vol. 33, No. 3.

[21] http://www.oclc.org/research/projects/pmwg/wg1.htm

[22] http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf

[23] http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

schema could be built, as well as a shared departure point for schema developed in different settings.

PREMIS working group

Release of the framework prompted new questions about preservation metadata and its use – questions such as what subset of information covered in the framework is *essential* for preserving digital materials over the long-term? How can this information be translated into *implementable* preservation metadata elements? How should preservation metadata be *created and maintained* in operational digital archiving systems?

To address these and other questions, OCLC and RLG sponsored a second working group: PREMIS (PREservation Metadata: Implementation Strategies)[24]. PREMIS was composed of more than thirty international experts in preservation metadata, drawn from libraries, museums, archives, government agencies, and the private sector. The working group's objectives were two-fold: first, using the framework as a starting point, to define a core set of implementable, broadly applicable preservation metadata elements, supported by a data dictionary offering guidelines and recommendations for populating and managing the elements; second, to identify and evaluate alternative strategies for encoding, storing, managing, and exchanging preservation metadata – in particular, the core elements – in the context of digital archiving systems.

In September 2004, PREMIS released *Implementing Preservation Repositories for Digital Materials: Current Practice and Emerging Trends in the Cultural Heritage Community*.[25] This report presents results from a survey addressing various aspects of existing and planned digital preservation repositories, including organizational mission, user communities, repository services, funding, characteristics of archived content, rights management policies, and of course, how metadata is being used to support repository processes, functions, and policies. Nearly 50 responses were received from institutions in 13 different countries; these institutions included libraries, archives, and museums, among others. Caution should be exercised in extrapolating the survey results to the entire digital preservation community – in particular, respondents were heavily skewed toward US libraries – but they nevertheless provide a valuable sampling of current approaches toward implementing preservation repositories. Survey responses underscored a number of issues impacting preservation metadata, including the extent to which repository architectures are informed by OAIS; the needs of repository stakeholders; methods for obtaining metadata for archived digital objects; types of metadata currently used by repositories; the nature and use of rights management metadata; access mechanisms for archived materials; and strategies for meeting long-term preservation objectives.

PREMIS followed up the survey report with the 237-page *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*, released in May 2005.[26] The report includes the PREMIS Data Dictionary 1.0, a comprehensive guide to

[24] http://www.oclc.org/research/projects/pmwg/

[25] http://www.oclc.org/research/projects/pmwg/surveyreport.pdf

[26] http://www.oclc.org/research/projects/pmwg/premis-final.pdf

core metadata needed to support long-term digital preservation. In addition to the Data Dictionary, the PREMIS final report includes a number of additional materials, including a report discussing special topics regarding the Data Dictionary; a glossary; and a set of examples illustrating use of the Data Dictionary for a variety of digital material types and digital preservation contexts. PREMIS also developed a set of XML schema[27] to support use of the Data Dictionary by institutions managing and exchanging PREMIS-conformant preservation metadata.

The Data Dictionary is organized around a data model consisting of five entities associated with digital preservation: *Intellectual Entity* (a coherent set of content that is described as a unit: e.g., a book); *Object* (a discrete unit of information in digital form: e.g., a PDF file); *Event* (a preservation action: e.g., ingest of the PDF file into the repository); *Agent* (person, organization, or software program associated with an Event: e.g., the publisher of the PDF file who deposits it into the repository); and *Rights* (one or more permissions pertaining to an Object: e.g., permission to make copies of the PDF file for preservation purposes). The Data Dictionary provides detailed descriptions of metadata associated with the Object, Event, Agent, and Rights entities, along with guidelines for implementation and use. Metadata for Intellectual Entities was considered out of scope, because it was felt that this information was already addressed in existing schema focusing on descriptive metadata.

A maintenance activity[28] has been set up to manage the current versions of the Data Dictionary and XML schema, and to coordinate future revisions. Currently, the maintenance activity takes the form of a Web presence hosted by the Library of Congress, but will soon be expanded to include leading institutions in digital preservation from around the world.

The work of PREMIS represents a significant step forward in terms of closing the gap between theory and practice in preservation metadata, and represents the only cross-institutional, cross-domain consensus-building activity in this area. Perhaps most significantly, the PREMIS Data Dictionary is based on the accumulated experiences of many institutions, representing a variety of domains but sharing a common need to set up and manage digital preservation capacity.[29] There is still much work to be done, especially in terms of testing the Data Dictionary in multiple domains and digital preservation contexts. Looking toward the future, widespread adoption of the Data Dictionary may help establish standardized practices for managing preservation metadata, enhance interoperability in a distributed network of digital repositories, and encourage potential economies from sharing and re-using certain forms of preservation metadata across repositories.

'The work of PREMIS represents a significant step forward in terms of closing the gap between theory and practice in preservation metadata…'

[27] http://www.loc.gov/standards/premis/schemas.html

[28] http://www.loc.gov/standards/premis/

[29] Although the membership of PREMIS was predominantly cultural heritage institutions, representatives from other domains contributed valuable perspectives as well.

* * *

In tracing the relatively brief history of preservation metadata, there is a discernible progression from concept to implementation, beginning with the high-level framework provided by OAIS, and culminating in the implementation-oriented PREMIS Data Dictionary. But significantly, both endpoints – OAIS and PREMIS – are anchored in consensus. Along the way, of course, there was a great deal of activity at the institution level aimed at resolving local challenges posed by preservation metadata, but in the end, drawing together the strands of fragmented effort into a framework of collaboration and consensus is the most potent strategy for developing collective solutions to shared problems in digital preservation.

Packaging metadata and content together in digital repository systems: METS

As the preceding discussion suggests, a digital object can accumulate a mushrooming quantity of metadata over time – not just preservation metadata, but also resource discovery, administrative, and other forms of metadata. This raises a critical question: how can all of this metadata be organized and linked to its associated content? Several solutions, essentially frameworks to package disparate metadata, have been proposed, including the Sharable Content Object Reference Model (SCORM)[30], the MPEG-21 Digital Item Declaration Language (DIDL)[31], and *METS* (Metadata Encoding and Transmission Standard)[32]. Of these, METS has the greatest potential for this purpose, as it was designed to implement the OAIS Reference Model's abstract model of an Information Package.

METS is an XML schema designed specifically as an overall framework within which all the metadata associated with a digital object can be stored. As such, it can function as either an OAIS SIP (Submission Information Package), DIP (Delivery Information Package), or, crucially here, as an AIP (Archival Information Package).

A METS file comprises four major constituent sections:

- a file inventory for all the files associated with the digital object (such as still image files, text, video or audio files)
- a section for administrative metadata (such as technical information about the files, rights management information, information on the source from which the object was made, and digital provenance information)
- a section for descriptive metadata (including bibliographic information and any other information on the intellectual content of the item necessary for users to find it and assess its value)
- a structural map, which indicates in a hierarchical manner how the various components of the item relate to each other, so allowing its constituent elements to be navigated by the user.

[30] http://www.adlnet.org/scorm/index.cfm

[31] http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm

[32] http://www.loc.gov/standards/mets/

These sections are linked to each other by means of identifiers: thus, an item in the structural map corresponding to a page in a digitized book may have pointers to the files in the file inventory which contain the scanned image of that page or a marked-up version of its constituent text, another pointer to the part of the descriptive metadata section which contains a full description of its intellectual content, and another to the part of the administrative metadata section which contains technical and rights information necessary to deliver the images or text.

METS allows two approaches to the storage of the metadata and data associated with a digital object: both may be either stored internally within the METS file, or held externally and referenced from within METS. The former offers the advantage of allowing everything associated with an object to be held (and archived) together, but can produce enormous files (particularly if the object includes image or video data). The latter produces more manageable files, but is vulnerable to the referenced objects being moved or removed from their stated locations. An overall architecture, which incorporates decisions on which method to use, needs to be drawn up at the beginning of a METS implementation.

‘METS allows two approaches to the storage of the metadata and data associated with a digital object: both may be either stored internally within the METS file, or held externally and referenced from within METS.’

The content of these sections is not prescribed by METS itself: any XML data or metadata (distinguished by differing XML namespaces) may be used. However, the METS editorial board does recommend a number of schemas[33], which will render the METS record more readily interchangeable: these include MODS (Metadata Object Description Schema)[34] and MARC-XML[35] for descriptive metadata, and MIX[36], METSRights[37] and TextMD[38] for administrative metadata. The set of PREMIS XML schema, in their current form, can also be used as METS extension schema; future work will look at the desirability of developing a more METS-specific PREMIS schema implementation.

The flexibility built into METS may cause problems in terms of interoperability. When such varied content, handled in such a variety of ways, is allowed within a METS file, it becomes more difficult to interchange METS records. This may be mitigated to some extent by the use of METS Profiles[39], XML files used to document the way in which METS is implemented within a project. These documents list, amongst other things, the content schemes used within a METS file, the system of identifiers, whether metadata is embedded or referenced, and how it is structured within the file. These do not allow the

[33] http://www.loc.gov/standards/mets/mets-extenders.html
[34] http://www.loc.gov/standards/mods/
[35] http://www.loc.gov/standards/marcxml/
[36] http://www.loc.gov/standards/mix/
[37] http://cosimo.stanford.edu/sdr/metsrights.xsd
[38] http://dlib.nyu.edu/METS/textmd.xsd
[39] http://www.loc.gov/standards/mets/mets-profiles.html

automatic transfer of METS files between systems, but are designed to help an implementer understand another body's usage of METS and how it can map to their own. The flexibility of METS implies that its practical implementation can be very flexible as well: any system capable of handling XML documents can be used to create, store and deliver METS-based metadata.

'The flexibility of METS implies that its practical implementation can be very flexible as well: any system capable of handling XML documents can be used to create, store and deliver METS-based metadata.'

A practical example of its implementation may give some indication of how it may be used in a complex environment. The Oxford Digital Library[40] at Oxford University aims to bring all of the university's libraries digitization projects under a single framework: these include collections newly scanned from materials held in its collections and legacy projects created within the university from previous digitization work. The range of materials is diverse, from medieval manuscripts to political cartoons, which necessitates a metadata scheme that can handle a wide variety of demands within a logical and extensible framework.

Newly created materials are handled by using a webform-based input mechanism, whose backend is a mySql database. Metadata is input following strict guidelines, which are applied to each new project, and additional, primarily administrative, metadata is derived from the data objects themselves and such sources as the database used to control workflow in the digitization studio. Once the items have been scanned and the metadata checked, a php script creates the METS file which becomes the sole metadata record for that object: mySql is therefore used only as one route toward the creation of the METS file, which then acts as the master record. METS files for legacy projects, which will conform to the same profile as that used for new material, are created in a variety of means, including XSLT transformations (for XML material), or by writing routines for the proprietary databases in which some are held to convert their contents straight into METS.

Implementing METS, as can be seen from this example, does not require any specialist skill sets beyond standard XML knowledge and experience: they can be created in an extensive number of ways, including generation from any software package that allows text output. Because METS files use XML as their architecture, they may readily be converted into almost any given output mechanism. The resources required for a METS-based system are, therefore, no more than one requires for handling any other form of XML object.

METS is at its strongest when dealing with a wide variety of materials which need to be handled flexibly but in an organized and coherent manner. Its structural metadata facilities and system of linkages make it particularly useful when dealing with

'METS is at its strongest when dealing with a wide variety of materials which need to be handled flexibly but in an organized and coherent manner.'

[40] http://www.odl.ox.ac.uk

items with a complicated internal structure, or those which incorporate complex webs of links. Materials with complex metadata requirements at all levels of their internal structure (such as an electronic journal which will require both monograph and analytic metadata) benefit particularly from METS's ability to handle metadata for any component at any level of granularity. It is also suited to simpler materials (such as a single image), as very little of the METS architecture is obligatory.

'Because it is non-proprietary, XML ensures that archival information is not tied to any given software package, and that it will not become obsolete as many such packages rapidly do.'

METS in many ways represents a useful solution to the requirements of digital preservation. It is written in XML, which (along with its predecessor SGML) has long been acknowledged as a robust and human-readable format for the archiving of metadata[41]. Because it is non-proprietary, XML ensures that archival information is not tied to any given software package, and that it will not become obsolete as many such packages rapidly do. Its flexibility ensures that archived metadata in XML should be readily usable in future deliverable mechanisms and that it should be interchangeable with other archives.

METS is designed to act as an Archival Information Package (AIP) as well as a medium for the submission and delivery of materials. OAIS's four categories of metadata (content information, preservation description information, packaging information and descriptive information) are either inherent in a METS file or can be incorporated into it: the descriptive metadata section, for instance, holds descriptive information, the file section content information, and the structural map packaging information. No conceptual leap is required to fit METS into the OAIS landscape.

Two factors, however, do need consideration before METS will fully comply with the OAIS model. Firstly, interoperability between a given archive and other OAIS-compliant repositories must be addressed by providing a METS profile to document the implementation of METS in the archival context: without such a profile, which must be referenced from the METS file itself, it will be much more difficult to "unpack" the archival object in the future. The importance of interoperability extends beyond METS to the PREMIS Data Dictionary as well (see the last paragraph of the next section for a discussion of this point). Secondly, the XML schemas or other metadata schemes used to record a METS file's component metadata need archiving as well, in order to prevent the file becoming invalid as these schemes are amended in the future.

Future directions

Most activity to date in the area of preservation metadata has been devoted to schema development; it is perhaps not too much to say that this activity culminated in the release of the PREMIS Data Dictionary in May 2005. If the Data Dictionary does become a standard in the community, a critical gap will have been filled, and preservation metadata activities can focus energy and resources on other problems. Areas of need that future

[41] See Coleman, James; Willis, Don *SGML as a Framework for Digital Preservation and Access* CLIR, 1997

preservation metadata work might address include automated tools, collaborative metadata management strategies, and exchange of archived content and metadata in a distributed network of repositories.

‘If the costs of preservation metadata are not to rise to prohibitive levels, automated tools must be substituted for human mediation in the workflow wherever possible.’

If the costs of preservation metadata are not to rise to prohibitive levels, automated tools must be substituted for human mediation in the workflow wherever possible. There is particular need for tools to extract and process required metadata from digital objects at the time of ingest into the repository. Some progress has already been made on this front. The NLNZ’s Preservation Metadata Extract Tool[42] harvests information from digital file headers, such as Microsoft Word, TIFF, WAV, and bitmaps, and outputs it in XML format. The JHOVE (JSTOR/Harvard Object Validation Environment)[43] automatically identifies, validates, and characterizes digital object formats, such as TIFF, PDF, and XML. Currently, much of the work relating to preservation metadata tools focuses on technical metadata (i.e., information having to do with the object’s format). Further work is needed to support other forms of preservation metadata; in addition, tools are needed that support formal preservation metadata schema like PREMIS. Once developed, preservation metadata tools need to be surfaced in a variety of digital asset management environments, like DSpace or Fedora.

Collaborative metadata management strategies can offer efficient, economical ways of acquiring and maintaining certain forms of preservation metadata – in particular, by leveraging opportunities for sharing and re-use, and diffusing metadata capture throughout the information lifecycle. Classes of digital objects sharing a common format, material type, origin, etc., will also share certain forms of metadata that apply to any object belonging to the class. A new object can therefore inherit a portion of its metadata from existing objects of the same class. This suggests opportunities for sharing and re-use of preservation metadata. The UK National Archives’ PRONOM File Format Registry[44] holds technical metadata about specific file formats, as well as descriptions of software needed to create, render, and migrate these formats. Repositories managing objects in these formats can point to metadata in the PRONOM registry, rather than creating and maintaining it locally. The metadata is created once, and then re-used many times.[45]

Collaborative metadata management can also improve efficiency in regard to the timing of metadata capture. Rather than waiting until the time of repository ingest to create and/or assemble all of the metadata necessary to support long-term preservation, it would

[42] http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction
[43] http://hul.harvard.edu/jhove/
[44] http://www.nationalarchives.gov.uk/pronom/
[45] Of course, the use of registries rather than locally stored metadata creates an external dependency that may violate the preservation policies of some repositories. In these circumstances, the repository may choose to store and maintain the metadata locally despite the economic advantages of the registry.

be more efficient (and likely cheaper) to collect metadata at the point in the object's lifecycle when it is most readily available. This of course requires some coordination across the various entities involved in an object's creation and management prior to repository ingest. Automatic Exposure[46], an initiative led by RLG, has opened a dialog with digital scanner and camera manufacturers to explore possibilities for automatic capture of the technical metadata enumerated in the NISO Z39.87 standard (Technical Metadata for Digital Still Images)[47] at the time a digital image is created. There is much more work to be done to exploit the benefits of collaborative metadata management, including determining what kinds of preservation metadata can best be managed through this approach, what forms of collaborations need to be arranged to meet community-wide objectives, and how these collaborations can be sustained over the long-term. The release of the PREMIS Data Dictionary should facilitate this work, since it can serve as a shared reference point for discussions occurring across multiple stakeholders and domains.

'…there is a need to explore the implications of exchanging preservation metadata across a network of heterogeneous digital archiving systems.'

Finally, there is a need to explore the implications of exchanging preservation metadata across a network of heterogeneous digital archiving systems. Most sources predict that future large-scale digital preservation efforts will be taken up by distributed networks of repositories (see, for example, background documentation for NDIIPP). In such an environment, it will be necessary to exchange archived content and metadata across the network, and ultimately, across digital archiving systems with vastly different architectures and technical implementations. Even systems using the same preservation metadata schema, such as PREMIS, will implement the schema in different ways, and adhere to different metadata management policies. Given this, research is needed that explores the preparation, transmission, receipt, unpackaging, and ingest of information packages – i.e., archived content and metadata – in multiple system environments. For example, one repository could prepare an information package for transfer to a second repository, which would accept the package through its ingest mechanisms, unpackage it, and incorporate the content and metadata into its own archiving system. These steps would then be repeated in reverse, with the same package transferred back from the second repository to the first. The information packages can then be analyzed to determine how and to what degree the preservation metadata was altered as it was transformed from stage to stage and moved from system to system. Work of this kind would provide insight into the effects of inter-repository transfer of preservation metadata as it moves into, through, and out of a series of heterogeneous digital archiving systems.

Conclusion[48]

There has been much accomplished in the area of preservation metadata, but even so fundamental questions, bearing on the scope and depth of the information needed to

[46] http://www.rlg.org/longterm/autotechmetadata.html

[47] http://www.niso.org/standards/standard_detail.cfm?std_id=731

[48] Parts of this section are adapted from Lavoie, B. (2004) "Preservation Metadata: Challenge, Collaboration, and Consensus" *Microform & Imaging Review* Vol. 33, No. 3.

support digital preservation, remain unsettled. This is largely because the digital preservation process itself remains unsettled – it is difficult to anticipate the metadata needed to support technical and administrative processes that are not fully developed, are not fully tested, and in some ways, are not even fully understood. Compounding the problem is the proviso that preservation metadata recommendations must be restrained by economic realities. Creating and maintaining metadata is expensive, so any recommended preservation metadata elements should be backed by persuasive evidence of necessity, as well as practical means for populating them.

Collaboration has proven to be an effective means to shape preservation metadata requirements within the bounds of these obstacles and constraints. Pooling expertise from a variety of institutional perspectives helps to mitigate the uncertainties associated with digital preservation. Similarities and differences across a range of digital preservation activities, exposed in the course of collaborative discussions, help draw the boundary between essential and non-essential preservation metadata, and ultimately lead to best practices that are both sensible and economical.

The art and science of digital preservation has advanced considerably in recent years, and institutions implementing a digital preservation capacity have a variety of resources available to inform and guide their work. Two of these resources are the PREMIS Data Dictionary – a consensus-based, comprehensive description of core preservation metadata, along with guidelines for its creation and maintenance – and METS – a means to establish the essential link between archived content and the metadata that, as noted earlier in the paper, makes the content self-documenting through time. Both of these resources will continue to be tested and refined as experience in using them accumulates; and hopefully, both will eventually form part of the permanent infrastructure needed to support sustainable, effective digital preservation programs.

> **‘Similarities and differences across a range of digital preservation activities, exposed in the course of collaborative discussions, help draw the boundary between essential and non-essential preservation metadata, and ultimately lead to best practices that are both sensible and economical.’**