

Using PREMIS to support preservation of digital assets at Yale

Yale University Integrated Access Council [2nd] Preservation Metadata Task Force 2006
October 2006

Member of the PREMIS Task Force

Matthew Beacom, Metadata Librarian, Library Catalog Dept. (Co-chair)
Reed Beaman, Associate Director for Biodiversity Informatics, Peabody Museum
Lee Faulkner, Media Director, Digital Media Center for the Arts
David Gewirtz, Project Manager, Library Projects, ITS
Kevin Glick, Electronic Records Archivist, Library Manuscripts and Archives
Rebekah Irwin, Catalog Librarian for Digital Projects, Beinecke Library (Co-chair)
Edward Kairiss, Director, Instructional Computing Instructional Technology, ITS
Daniel Lee, E-Publishing/Internet Marketing Manager, Yale University Press
Youn Noh, Digital Resources Catalog Librarian, Library Catalog Dept.
George Ouellette, Senior Programmer Analyst, Library ILTS
Thomas Raich, Associate Director, Information Technology, Art Gallery
David Walls, Preservation Librarian, Library Preservation Dept.

Background

In March 2006, the Yale University Library Integrated Access Council accepted the recommendation of its first Preservation Metadata Task Force to adopt the PREMIS model as the basis for preservation metadata for its digital information assets.¹ As a practical next step toward implementation of PREMIS-based preservation metadata, a second task force was charged in April 2006 to investigate, develop, and propose a metadata element set and usage guidelines based on the PREMIS model.

The scope of PREMIS is limited to metadata that supports preservation activities. PREMIS-based metadata is only one part of a larger package of metadata needed to support the use and re-use of digital assets held within a repository. Other metadata standards complement PREMIS. Additional metadata for digital information objects would include, for example, MARCXML or MODS for descriptive metadata that supports discovery of information objects, NISO MIX for technical metadata that describes the physical characteristics of images, and METS to contain or wrap together combinations of metadata types for transmission between applications or repositories. PREMIS is part of a set of metadata tools with which Yale can manage its digital assets, provide for retention and re-use for the long term, and build information services for the Yale community and the world.

In our discussions and analysis of using PREMIS at Yale, two of our most critical concerns have been effectiveness and cost. To be of value, the proposed element set and guidelines must both support the preservation of digital information assets at Yale and meet the practical needs of digital asset administrators. Administrators need to weigh the cost and effectiveness of generating and storing additional technical and

¹ [Recommendation to Adopt PREMIS as a Preservation Metadata Model for the Yale University Library \(1/12/06\)](http://www.library.yale.edu/iac/documents/PMTF_Rec_Adopt_PREMIS_v32.pdf) [PDF, 11 p.]

http://www.library.yale.edu/iac/documents/PMTF_Rec_Adopt_PREMIS_v32.pdf

preservation metadata, taking into account the stability of the format, the value of the object, and obligations to third parties.

PREMIS is designed to be an effective and inexpensive—“implementable”—tool that provides the metadata or information needed to preserve digital information assets for the long term.² To that end, the PREMIS elements are designed to be information that is collected through automated or semi-automated processes. No PREMIS element should require manual input per digital object. Although they may not be fully automated across every application, they can certainly be semi-automated. The process to create or capture preservation metadata should be economically sustainable.

Costs also may arise from developing the automated tools needed to create or capture preservation metadata and the policies to support their use. For example, the base profile has only a tiny number of elements that are required, but each element requires work to develop tools and policies. Yale would need to do this development quickly and efficiently, or the result could be a protracted development phase.

Measuring the effectiveness of PREMIS metadata needs to be based on experience with the performance of the metadata in actual instances of use. Preservation of digital information assets is new. Trials using PREMIS are now underway around the world. Two of our recommendations below address the need to engage with peers implementing and evaluating PREMIS while beginning to use PREMIS locally in an experimental—evaluative—spirit.

There is, as well, a third related and critical issue that must be considered with effectiveness and cost. That concern may be thought of as *relevance*. In short, all digital assets at Yale may not warrant an application of metadata suitable for long-term preservation. Although administrators of digital asset collections need guidelines for application and use of preservation metadata at Yale, these guidelines must be relevant to a wide range of digital asset collections. For example, administrators of e-records in a university archive may have a different obligation for long term preservation than do administrators of digital surrogates of items in museum and library collections, or administrators of e-reserve course materials. The range of digital asset collections at Yale requires an adaptable framework for determining the appropriate degree of preservation responsibility per collection, repository, or administrative unit.

PREMIS cannot make policy decisions. PREMIS cannot provide guidance on appropriate levels of responsibility for preservation of Yale's digital assets. However, PREMIS can supply a critical piece of the digital preservation infrastructure at Yale. It is a building block well-suited to Yale's wide ranging digital asset collections and systems implementations. What is needed is a larger policy framework for applying the PREMIS model at Yale.

² [Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group](http://www.oclc.org/research/projects/pmwg/premis-final.pdf) [PDF:3.2MB / 237p.] <http://www.oclc.org/research/projects/pmwg/premis-final.pdf>

Recognizing this diversity of needs among digital asset administrators, we understand that there is no “one size fits all” approach to preservation metadata. The library and its peers at Yale need to develop policies to determine when to apply PREMIS. The task force recommends that Yale develop a small set of preservation metadata profiles for distinct yet relatively generic “levels” of obligation for preservation of digital assets. The profiles would guide appropriate use of PREMIS across a continuum of obligation. Repository administrators would review semantic units in PREMIS to insure that they are capturing the information needed to extend a base profile or abbreviate a full profile to their domain-specific profiles.

Between the extremes of a base profile or a full profile, the library and its peers at Yale need to develop a set of profiles that relate preservation policies with appropriate metadata requirements. Development of such profiles will depend on a number of policy decisions made by various organizational units at Yale, such as the library, the gallery, natural history museum, and ITS. The staff developing these profiles should be experienced in preservation activities, digital information technology, and metadata services. These profiles should be generic or neutral with respect to specific information systems implementation. Two levels are immediately apparent, a base level and a full level.

Two Profiles

Base profile. This profile would provide guidance to creators and aggregators of digital information assets on what elements they should generate or capture as they create or ingest digital information assets. The base profile could thus be used across many digital asset collections at Yale.

The base profile, proposed in this report, is a sub-set of the full PREMIS profile that is the main body of this report, and it is intended to be temporary until the library and its peers at Yale have developed digital preservation policies. The Base Profile is presented here.

The base profile is intended to apply to objects of all formats. The base profile provides a means of identifying and tracking objects. Small errors and inconsistencies can easily make these functions nontrivial.³ In order for a repository to scale, basic tracking mechanisms need to work. The base profile may be supplemented incrementally with global updates once Yale-wide preservation policies have been established and communities within the university have developed their own domain-specific profiles.

³ Shirky, C. (2005, December). AIHT: conceptual issues from practical tests. *D-Lib Magazine*, 11 (2). Retrieved September 29, 2006, from <http://www.dlib.org/dlib/december05/shirky/12shirky.html>

PREMIS Base Profile

- 1. **objectIdentifier**
 - 1. 1. objectIdentifierType
 - 1. 2. objectIdentifierValue

- 2. **objectCharacteristics**

- 2. 1. **fixity**
 - 2. 1. 1. messageDigestAlgorithm
 - 2. 1. 2. messageDigest
- 2. 2. **format**
 - 2. 2.1. formatDesignation
 - 2.2.1a. formatName
 - 2.2.1b. formatVersion

The **Object Characteristics** category records technical properties such as size and format that are applicable to all or most digital formats.

In order to insure the **fixity** of a digital object, i.e. to certify that an object has not been altered in an unauthorized way, two **Message Digests**, or fixity checks, should be performed automatically by the repository upon ingest and at a later time. The **Message Digest Algorithm** defines the specific algorithm used. The Message Digest is the output value that can be stored and compared in future fixity checks.

An **Object Identifier** should be automatically created by the repository system at the time of ingest. The object identifier serves as the repository's primary identifier in order to insure that identifiers are unique and usable by the repository. Until a repository for long-term preservation is established, creators should assign their own unique identifiers to digital materials intended for long-term preservation.

The **Object Identifier Type** designates the domain within which the object identifier is unique.

The **Object Identifier Value** designates the value of the objectIdentifier.

The **Format and Format Designation** categories accurately and automatically identify file formats upon ingest into a repository.

Format Name designates the file format, e.g. "Adobe PDF"

Format Version records the version of the format named in **formatName**, e.g. "6.0"

Full profile. This profile would provide guidance to administrators of digital information assets acting as trusted custodians of material deemed to be of long-term value. The full profile is a draft that needs to be fine-tuned through experience with actual instances of use at Yale. We can only go so far based on documents, theories, and the experience others are beginning to have. Experience using PREMIS will determine which elements in the PREMIS model are necessary at Yale. For an illustration of how the full profile may be applied to a particular thing, see the accompanying example that is based on material from the digitization of slide images from the Classics Dept.

The full profile is represented in the accompanying spreadsheet that describes the PREMIS model element by element, providing rationale, usage guidelines, and estimates of obligation for each element.

Recommendations

Taking action on preservation metadata at Yale will be complex because Yale is a complex organization. The following recommendations call for developing a coherent, sustained effort to preserve digital information assets at Yale by beginning (largely) with the library. Yale can use its library to prototype solutions for policy, technology and processes. The actions for the library named below should be undertaken as the first step in a larger process to successfully guide and support the creation, acquisition, use and preservation of digital information assets at Yale.

- Experiment with and evaluate PREMIS metadata in use at Yale.
 - Action: Include PREMIS as part of the Yale University Library VITAL Fedora implementation and to make preliminary report in June 2007 to update IAC on progress.
- Participate in the PREMIS Implementers' Group to work collaboratively with other PREMIS implementers, such as the Library of Congress, the National Library of Australia and Stanford University.
 - Action: Assign staff involved with the VITAL/Fedora implementation of PREMIS and the library's Digital Preservation Committee to participate in the PREMIS Implementers' Group. The implementation group will update IAC in June 2007 on the status of Yale's participation.
- Establish a coherent policy framework and governance structures for the preservation of digital information assets at Yale.
 - Action: Digital Preservation Committee (DPC) to develop digital preservation policies for Yale University Library and to make a preliminary report to IAC in June 2007
 - Action: Digital Landscape Committee to develop digital preservation policies for Yale University.
 - Action: IAC and Yale University Library to support use of the base preservation metadata profile throughout Yale by providing critical services to users.
 - Objective: establish persistent naming services for digital assets in Library by June 2007.
 - Objective: establish validation services for digital assets in Library by June 2007.
 - Objective: establish fixity check services for digital assets in Library by June 2007.

Guide to the full profile in the accompanying spreadsheet

The PREMIS data dictionary has 22 metadata semantic units or data elements (19 contain nested sub-elements) divided across four types of entities: Object, Event, Agent, and Rights. The body of the evaluation addresses each element and nested sub-element in the PREMIS data dictionary.

The full preservation metadata profile is contained in an Excel spreadsheet. All the PREMIS semantic units are broken out as individual elements and sub-units across the four types of PREMIS entities: Object, Event, Agent, Rights. For each element and sub-elements, we suggest a responsible agent, indicate whether the data is likely to be automatically or manually supplied, provide a rationale for the use of the element and, give usage guidelines, and note any applicable data constraints on the value for each element or sub-element. Additionally, we note the applicable object category to which the element can be applied and state the level at which the element is nested within the matrix of PREMIS semantic units. Each of these aspects is explained in more detail below.

Semantic Unit: the PREMIS name for the elements used in this evaluation.

Obligation (PREMIS): the level of obligation assigned in the PREMIS Data Dictionary. Possible values are Mandatory and Optional.

Obligation (Yale Full Profile): the level of obligation assigned by this group for the full profile. These assignments of obligation are our current best guess and are likely to be adjusted after further experience is gained with use. Possible values are Mandatory, Recommended, Optional, and Not Recommended.

Responsibility: an indication of who is obliged to provide the metadata. The possible choices are the *creator* of the digital asset, the *repository* that is accepting the digital asset, or the *submitter*.

For some elements, more than one responsible party may be designated. Flexibility in designating responsibility is required for at least the following reasons: (1) parties may assume joint responsibility for the generation of certain elements, (2) a repository may assume responsibility for metadata not provided by the creator or submitter if the object is of high value, or (3) designation of responsibility may depend upon relationships among the parties (e.g., if the repository is a client of the creator).

Generation: an indication of how the data for this element can be supplied; the only choices indicated here are *automated*, *semi-automated*, or *manual*. *Semi-automated* processes include use of default settings, batch processes, and templates.

Rationale: a brief statement declaring the value of the element to the repository (what functionality the repository gains by knowing this information).

Usage notes: Notes on best practices for using the element. This may be a rich target for comments as usage may depend greatly on the goals of the administrative unit and the implemented environment of any given repository.

Data Constraint: Any constraint on the data, such as integer, controlled vocabulary, container, none.

Repeatability: Possible values are repeatable or not repeatable. The value applies to all object categories to which the element is applicable unless stated otherwise.

Applicable Object Categories: Indicates which PREMIS model object entities the element can be applied to (Representation, File, and Bitstream). Each PREMIS semantic unit has defined object categories. Some may be applied to all, other to just one of the object categories.

Parent: The parent element named is the container closest to the element being described.