IFLA
2005
OSLO

**World Library and Information Congress:
71th IFLA General Conference and Council**

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

08 June, 2005

Code Number:	**060-E**
Meeting:	**150 SI - ICABS (IFLA/CDNL Alliance for Bibliographic Standards)**

Preservation Metadata Standards for Digital Resources: What we have and What we need

Sally H. McCallum
Library of Congress
Washington DC, USA

Abstract:

A key component for the successful preservation of digital resources is going to be the metadata that enables automated preservation processes to take place. The number of digital items will preclude human handling and the fact that these resources are electronic makes them logical for computer driven preservation activities. Over the last decade there have been a number of digital repository experiments that took different approaches, developed and used different data models, and generally moved our understanding forward. This paper reports on a recent initiative, PREMIS, that builds upon concepts and experience to date. It merits careful testing to see if the metadata identified can be used generally and become a foundation for more detailed metadata. And how much more will be needed for preservation activities? Initiatives for additional technical metadata and document format registries are also discussed.

Core Metadata for Preservation: PREMIS

Origins

The Preservation Metadata Implementation Strategies (PREMIS) project grew out of the experience of the last decade.(1) There has been significant work on repository systems in the library community and particularly among the component institutions of ICABS and their collaborators. This work inevitably involved the design of some type of formal or informal data models and the identification of data elements for the preservation function, even though it often had broader goals than preservation, being focused on access and distribution issues. Some of those projects were the Networked European Deposit Library (NEDLIB) project led by the National Library of the Netherlands and the Bibliothèque Nationale de France, CURL Exemplars in Digital Libraries (CEDARS) project from the United Kingdom, the Pandora project of the National Library of Australia, and various institutional initiatives such as those undertaken by OCLC, the National Digital Library experience of the Library of Congress and others.

Interestingly, all of these projects addressed at some point their relationship to the Open Archival Information System (OAIS) reference model (2), which was first articulated for space data systems and later became an ISO standard (ISO 14721). The OAIS model has had a unifying impact on the investigations over the last decade if only to provide a language at a high level to support discussions. The Archival, Submission, and Dissemination Information Packages (AIP, SIP, and DIP, respectively) as basic conceptual components in the implementation of digital repositories are commonly understood. These information packages are made up of four parts related to the information object that is being treated: the content information, packaging information, description information, and, our focus, preservation information. In 2002, a project sponsored by OCLC and RLG did an excellent job of bringing together in one framework the models and metadata specified in the above projects and fitting them into the broad concepts of the OAIS reference model.(3) The PREMIS working group's primary task was therefore to pick up those threads and translate them into a set of implementable data elements, via a data dictionary.

Goals

The PREMIS project was a multi year working group endeavor with participation from institutions with major implementations worldwide. Representatives from Australia, New Zealand, the United States, Great Britain, Netherlands, and Germany contributed in various ways, some rising at early hours to participate in weekly conference calls. Work planned for one year took two but the result is a highly refined set of elements that can serve as a foundation for implementations.

The effort had several related goals, all practical and intended to give concepts an implementation foundation. The original goals included identification of a "core" set of metadata and development of a data dictionary for that metadata, both of which are now successfully completed. Experimenting with the data dictionary will be the best method for articulating alternative strategies for implementation, the third goal. The final goals, pilot tests of the data dictionary and cooperative programs based on the core elements are to follow the current work.

Survey

Work began by surveying a number of implementations of digital project repositories to identify current practices and trends for digital projects. The survey had 48 responses from 13 countries, a good rate for a developing area. The general conclusions from the survey (4), which served to inform the data dictionary work that proceeded in parallel and followed the survey, may be summarized as follows:

- There is widespread use of the OAIS reference model for framework and starting point for repository design.
- It is common practice to store metadata redundantly in repository systems; in an XML or relational database for rapid retrieval and flexible reporting and with the content object itself for self defining and preservation futures.
- There is extensive use of the Metadata Encoding and Transmission Standard (METS) for encoding the broader spectrum of metadata needed for digital objects, including preservation metadata; with MIX (Metadata for Images in XML) used within METS for technical image metadata.
- The current trend is to keep the original and also store several normalized and/or migrated versions of the content object, each with related metadata.
- Use of multiple strategies, even within an institution, is not uncommon in such an experimental and developing area.

In addition the survey showed that a number of distinctions were made for metadata relating to different types of objects (bit streams, files, collections, logical objects, etc.) and information indicating relationships between objects was frequently recorded. While a survey instrument in an emerging area like this is not definitive, the results were both interesting and useful in the data dictionary work.

Data Dictionary

Drawing on the earlier framework project (and indirectly from the several major projects of the last decade) and the information from the survey of digital repositories, the core elements data dictionary was then developed by the PREMIS working group.(5) Several decisions were made in the early stages of the project which are important for its practicality.

Core data elements were interpreted by the working group to mean "things most working preservation repositories are likely to need to know in order to support digital preservation." (6) The group intentionally did not treat some well-known aspects of preservation, such as detailed technical metadata for different media. Only technical metadata that was generally applicable across file formats was pursued by the working group for PREMIS.

Another important consideration embraced by the working group was that the metadata specified must be able to be automatically supplied and used, in so far as possible. This led to a preference for values from authorized lists over textual descriptions. It also relates to the working group's intent to make the data dictionary implementation independent. As the survey showed, repositories are already in production and for those in the planning stage, the systems environment in which they will reside may have special characteristics. The PREMIS core elements that are to be available to the repository are not necessarily explicitly stored in it. The elements could be stored in auxiliary systems, could be implicit in the business rules used by the repository, or could be stored within a local database or format. The important point is that the core data be available for conversion to some standard in the event of interchange. Or that the data be predictably available to any software that the repository might choose that expects the PREMIS core data to be accessible. Systems do not need to be reimplemented or specially designed in order to hold the PREMIS core in some standard format. This led the working group to define "semantic units" in the data dictionary instead of "metadata elements".

Data model

While this paper is too short for a detailed description of the data model, a few features are important to especially note. (The model is fully and well explained in the PREMIS report, see reference (5))

The model is *simple*. There are only 5 types of entities: *Objects, Events, Agents, Rights,* and the *Intellectual Entity* itself. The core-ness of the information that was included in the data dictionary was carefully observed. Thus, for example, descriptive metadata describing the Intellectual Entity, which may be a book, map, web site, etc., is left to the many standards such as MARC, MODS (Metadata Object Description Standard), and DC (Dublin Core) that already exist. Likewise detailed data about Agents is left to MARC, MADS (Metadata Authority Description Standard), vCard and other standards. Rights data is confined to that pertaining to permissions for preservation activities, since rights associated with access or distribution of the Object are not core to preservation activities. Detailed technical metadata and media and hardware documentation are not included but left to format experts to specify.

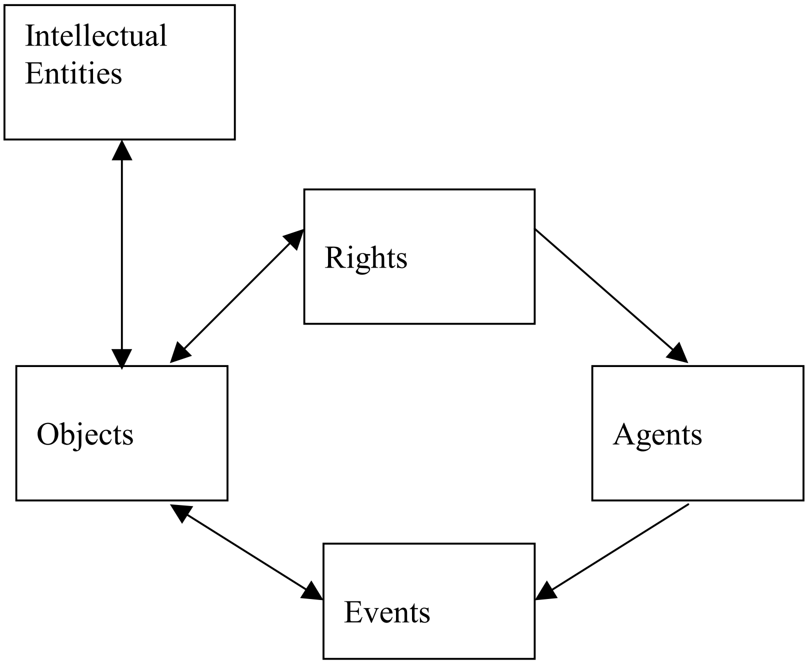

Figure 1: Basic PREMIS data model

Semantic units for Objects, the focal concept in the model, can be specified at three levels, providing flexibility to include information at the levels appropriate for the material and for the operation of a repository. These levels are *bitstream*, which is a component of the next level, the *file* (or *file stream*). A set of files needed for a complete rendition of an Intellectual Entity constitutes the highest level, the *representatio*n.

The Event entity which documents actions related to the Object is an important part of the model. A large variety of actions affect the preservation of digital material including modification of the Object, validity and integrity checks carried out, even requests for dissemination or reports. Events are also frequently related to relationships, since a derivation Event produces another Object and the relationship between the Objects is usually important to record for preservation purposes. The data dictionary provides several relationship semantic units related to record derivative and structural relationship information, dependencies, and other relationships.

An important aspect of the data model is what the working group called the 1:1 principle. New Objects created from existing Objects (copies, versions, transformations, etc.) are treated as new Objects and linked to the "old" Object by Event and relationship information. One of the findings in the survey was that repositories often keep multiple copies of an Object, and for preservation purposes, the data about each Object is important to be complete. Therefore relationship information provides the link without diminishing or making complicated the full recording of preservation information about the derivation. While, internally a repository may build data trees to reduce data redundancy, for interchange the repository needs to be able to forward an independent Object with full preservation metadata.

Next: Testing

PREMIS has been a carefully scoped, international collaboration that produced a data dictionary of metadata with the potential for enabling standard exchange of preservation information with digital material from electronic archives. It does not force specific architectures on the repositories but provides guidance for core preservation metadata. The PREMIS project, though global in participation, was sponsored by OCLC and RLG, and the Library of Congress has taken on the responsibility of the official web site for the next phase (7). All project documents and news can be obtained via that site.

The final goals of the project, a data dictionary test bed and cooperation built around the metadata can now be planned. An XML schema has recently been written for the semantic units identified in the data dictionary.(8) It needs to be used and tested by new projects and for exchange. However, it is hoped that existing implementations of repositories or projects planned with special architectures will also participate in the test bed, by analyzing their metadata, implicit and explicit, against the semantic units of the data dictionary. Meanwhile the data dictionary and the XML schema will be kept stable but subject to maintenance revision as experience is gained in the test bed.

Other Parts of the Puzzle

As noted above, there are other parts to the preservation metadata needs for digital media that were not defined by the PREMIS working group B for example, extensive rights metadata and detailed technical metadata, including digital format information.

Rights metadata

Rights metadata was narrowly defined for PREMIS, and one could argue that some access and dissemination information is important for preservation purposes. A number of initiatives, however, are addressing the rights expression language and messaging standards issues as related to access and dissemination. The Indecs work of the European Union, ONIX efforts from the publisher groups, and the Electronic Rights Management Initiative (ERMI) of the Digital Library Federation (DLF) are a few major investigations.

Technical Metadata

The PREMIS survey found that many repositories were using METS to bundle their digital object metadata, and that there was variety in the type and amount of technical metadata held, depending on what the repository could automatically collect. The one area where standards work had made significant progress is with metadata for image resources. A standard data dictionary was completed in NISO for trial use on 2002.(9) MIX, the METS extension schema based on the NISO data dictionary, is already, however, widely used.(10) The fast up-take of this standard and schema indicate that repositories are very interested in standards and guidance for detailed technical information. For detailed technical metadata the library community needs to collaborate with, or at least take careful note of, any emerging industry standards, as this level of metadata needs to be derivable from the objects B even more so than the PREMIS level information. The METS web site points to several locally developed technical metadata schemas for various types of material that can perhaps serve as the starting point for broader efforts to develop standards comparable to that for image data.(11)

Format Registries

A second, potentially valuable piece of the preservation metadata suite is easy access to electronic data format specifics. This information can sometimes be found at the web sites of the companies responsible for various data formats, if such a site exists, but this is not an efficient way to obtain the information. From the preservation perspective, knowledge of data formats assists in validation of digital objects at ingest or for integrity checks, it helps evaluate risk associated with various digital formats, and it indicates appropriate migration pathways for digital objects. An understanding of the file format can also help determine metadata that might be extractable from the digital object, thus helping to populate PREMIS and detailed technical metadata databases.

There have been two prominent projects to develop continually updated collaborative directories but it is not yet clear whether they can be sustained. One project is PRONOM, from the National Archives in the United Kingdom.(12) This registry started as a locally compiled tool needed by the National Archives to help combat software obsolescence by guiding the migration of documents. It was made web accessible in 2004 and in 2005 a greatly enhanced new release was made. With an emphasis on public records, this registry has been especially strong on text oriented software formats.

A second project which has progressed to a proof-of-concept stage is the Global Digital Format Registry (GDFR) that emerged from a DLF-sponsored meeting in 2003.(13) As soon as the model of this registry was released by staff at Harvard, a prototype format service was developed at the University of Pennsylvania, called the Format Registry Demonstration (FRED).(14) Through FRED, repository developers can experiment with how such a service might be useful, what services should it offer, how it could be maintained, etc.

This is an area that is not glamorous but appears to be important for preservation across all media B and a collaborative registry would be efficient for the community.

Conclusion

Step by step, building on past conceptual models and experiences in implementation, guidelines and standards are emerging for metadata in support of repository preservation activities. Repository builders no longer need to start from a "blank sheet of paper". Testing of the PREMIS core elements, attention to the detailed technical requirements, and collaboration on a data format registry are today's agenda for future developments.

References

(1) PREMIS official web site: http://www.loc.gov/standards/premis
(2) *Reference Model for an Open Archival Information System (OAIS)*. Washington, DC: Consultative Committee for Space Data Systems, 2002. (http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf)
(3) *A Metadata Framework to Support the Preservation of Digital Objects*. Dublin, Ohio: OCLC Online Computer Library Center, 2002. (http://www.oclc.org/research/projects/pmwg/pm_framework.pdf)
(4) *Implementing Preservation Repositories for Digital Materials: Current Practice and Emerging Trends in the Cultural Heritage Community*. Dublin, Ohio: OCLC Online Computer Library Center, 2004. (http://www.oclc.org/research/projects/pmwg/surveyreport.pdf)
(5) *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group, May 2005*. (http://www.oclc.org/research/projects/pmwg/premis-final.pdf)
(6) Ibid., p.ix.
(7) The official PREMIS web site is: www.loc.gov/premis/
(8) The PREMIS schemas may be obtained from http://www.loc.gov/standards.premis/schemas.html
(9) *Data Dictionary B Technical Metadata for Digital Still Images, NISO Z39.87-2002/AIIM 20-2002*. (http://www.niso.org/standards/resources/z39_87_trial_use.pdf)
(10) MIX may be obtained from http://www.loc.go/mix
(11) See http://www.loc.gov/mets
(12) For more information: http://www.nationalarchives.gov.uk/pronom/
(13) For more information: http://hul.harvard.edu/gdfr/
(14) For more information: http://tom.library.upenn.edu/fred/