Evaluating PREMIS in an Academic Research Library

Ivey Glendon and Gretchen Gueguen

Introduction

In December 2011, a group of digital preservation stakeholders at the University of Virginia (UVa) Library formed a task force to discuss preservation metadata for digital collections, and, specifically, the viability of applying the PREMIS (PREservation Metadata: Implementation Strategies) metadata standard to digital objects in the digital repository at UVa. Developed by an international working group from 2003 to 2005 and updated in 2008, PREMIS is a nationally recognized preservation metadata standard.[1] The Library of Congress maintains the PREMIS XML schema and also sponsors the PREMIS Maintenance Activity group that administers and provides updates to the standard. With PREMIS available as a national preservation metadata standard, staff from several departments at the library—including Metadata Management Services, Digital Curation Services, Preservation, and Special Collections—met for approximately six weeks to evaluate PREMIS and to formally convey to library management staff recommendations for levels of PREMIS implementation at the UVa Library.

This paper provides a survey of the collections at the UVa Library, a review of the PREMIS metadata standard, and an overview of the task force's findings related to evaluation of the standard and recommendations for PREMIS implementation at the UVa Library. The authors hope that by providing this example of one library's attempt to adopt PREMIS, other libraries can plan their own evaluation and implementation.

The University of Virginia Library: A History of Innovation

The University Library has always been central to the University of Virginia—Thomas Jefferson himself selected the books for the first library and placed it in the University's main building, The Rotunda.

The University of Virginia Library now comprises 11 individual libraries, in addition to "sister" libraries in the Law, Business, and Health schools with separate administrative structures. The Library holds more than 4 million print volumes, 16 million manuscripts and archives, 3 million microforms, and 7 thousand print subscriptions. The campus community includes nearly 1,000 faculty members, 20,000 undergraduate and 6,000 graduate students.

UVa strives to be a leader in innovation, particularly in the area of library technology. The first digitization projects at UVa began in the 1990s, and the library was an early participant in the development of the Fedora digital repository software. UVa developed Blacklight, an open-source OPAC, which led to the development of Hydra, a stack of technologies for repository development. Hydra is now supported by a large community of library investors and continues to grow.

By 2011, the UVa Library had digitized more than 15 million objects and had begun to develop workflows related to the accession and management of born-digital materials. The library had worked for many years on the development of Fedora and Hydra, and was beginning to plan for centralized manage-

Ivey Glendon is Metadata Librarian at the University of Virginia Library, e-mail: img7u@virginia.edu; Gretchen Gueguen is Digital Archivist at the University of Virginia Library, e-mail: gmg2n@virginia.edu

ment of digital objects through these tools. PREMIS could potentially standardize needed metadata for long-term management in this new environment. However, the library's organizational structure dictates that several organizational units share "management" of digital assets:

- Metadata Management Services (MMS): responsible for cataloging of the library's general collection as well as the usage and implementation of various metadata schema for digital projects
- Collection units: Many collecting units have material that has been digitized or is born-digital. These units (such as Special Collections or Arts and Media Services) perform the original description and arrangement of materials and are liaisons for the digitization process
- Digital Curation Services (DCS): Digitization services as well as management of long-term storage solutions
- Online Library Environment (OLE): Developers of VIRGO (Fedora/Solr/Blacklight, but not Hydra) and LIBRA (Hydra-based) – both the repository development as well as the web-enabled access to that repository
- APTrust and DPN: Two related projects headed by UVa related to long-term preservation and storage
- Preservation: The preservation department at the UVa libraries is largely responsible for analog preservation, however, their role in the digitization of materials for preservation purposes serves as a link between the analog and digital environment.

Because of the decentralized nature of organizational units at the Library, decisions related to metadata and digital preservation are necessarily assessed and considered for implementation across multiple service units. This necessitated the creation of a task for to address whether PREMIS would be a workable solution for all involved.

The PREMIS Task Force

In December 2011, the heads of Metadata Management Services, Preservation, and Digital Curation Services assembled the PREMIS Task Force (PTF) to examine the PREMIS metadata standard and its fitness for implementation in the library's repository. The task force included representatives from MMS, DCS, Special Collections, Arts and Media Services, and Preservation; while not formally represented, staff from the library's OLE group provided expertise as needed. The group would submit a final report to the management team to summarize the task force's findings and recommendations.

The PREMIS Standard

The first objective of the PTF in examining and evaluating PREMIS was to gain a basic understanding of what PREMIS is meant to do. PREMIS is a suite resources related to digital preservation: a data dictionary, a data model, and an XML schema. PREMIS contains semantic units, laid out in the data dictionary, rather than metadata elements and these units map precisely to the metadata elements in the PREMIS XML schema. The data model describes the five core entities about which information on digital preservation activities can be recorded:

- Intellectual entities
- Objects
- Events
- Agents
- Rights

An intellectual entity contains bibliographic provenance information for an item, such as a book or a map. The remaining four entities of the PREMIS data model capture information related to the digital preservation of digital assets related to that item. The object entity describes the basic characteristics of the digital object is that is stored in the repository. Object characteristics may include information about the object's creation, its size and format, or its unique identifier. Event entities gather information about actions that happen to an object or objects, such as creation or migration, including the date and time of the event, or a unique identifier associated with the event. Agent entities are actors—people, software, or organizations—that perform activities on an object. The rights entity contains information about permissions related to storing objects in a repository, such as a copyright statement for an object or access permissions and restrictions for a given object.[2] The Library of Congress working group expects and intends PREMIS to be represented in XML and has, to that end, developed an XML schema for use in PREMIS data exchange.

With this understanding, the PTF turned its attention to how PREMIS might serve digital preservation activities at the UVa Library. The PTF considered in its research and deliberations the holistic picture of the library, including collections, staff and tools for use in implementing a PREMIS-conformant digital preservation program.

The PTF's Findings

The PTF submitted a formal report summarizing their findings in February 2012. Following a basic evaluation of the schema and its fitness in tracking necessary preservation metadata, the group described the following:

- Types of resources that would benefit from tracking preservation metadata
- The role of PREMIS in digital preservation audit trails of archival digital materials
- Technical requirements tools available for implementation
- Recommendations for level of implementation
- Impact on staff workflows.
- The major conclusions of the report are summarized below.

Types of Resources that Would Benefit from Tracking Preservation Metadata

The PTF found that the types of resources that would benefit from tracking preservation metadata are those digital resources that the institution cares about preserving over time—that is, unique master digital files that were either created in a digital format or were digitized by the institution. As examples, the PTF examined three concrete examples of resources that currently benefit or would benefit from tracking preservation metadata and noted the specific types of PREMIS data that each project already captures or would seek to capture.

The first collection the PTF identified as a resource that would benefit from tracking preservation metadata is the library's archival audio and video materials. Since the materials reside on a diverse range of obsolete media formats, these materials are a high priority for digital conversion. Implementing PREMIS permits the capture and retention of detailed information about the digitization process and provides a place to store format-specific technical metadata, information about the software creation environment, as well as authenticity and fixity information. PREMIS events fully document the initial creation of digital preservation masters, as well as information about the analog source materials from which these files derive. PREMIS would also prove useful for documenting the creation of derivative access files from new digital preservation masters.

The second collection the PTF assessed was the output of the acquisition and processing of born-digital archives located in Special Collections. The collection contains more than 1500 disks received over the past 20 years. The volume of these digital holdings is expected to increase substantially in the coming years. Successful transfer of the content of these disks from their native environment into secure long-term storage requires the tracking of a substantial amount of technical metadata. PREMIS has been designed to capture this vital information such as checksum and object characteristics and tracks preservation events such as virus check and creation of a derivative.

The third collection evaluated is the ever-growing collection of digital surrogates of rare and unique materials created by Digital Curation Services. The unit has been using a home-grown application called Tracksys to manage digital production and automated workflows. In the process of performing its tasks, Tracksys captures metadata appropriate for retaining as PREMIS metadata. Production masters are eventually moved to StorNext, a high-performance enterprise data management tape storage solution provided by the University of Virginia Information Technology Services. Access derivatives are available through UVa's Blacklight-driven VIRGO interface and managed through its Fedora repository.

The Role of PREMIS in Digital Preservation Audit Trails of Archival Digital Materials

Having established that the library has and will continue to have content primed for PREMIS implementation, the PTF examined the role PREMIS might have in the library's digital preservation activities. The working group established that preservation metadata pertaining to all objects in archival storage should be stored in the PREMIS format in order to establish consistent and shareable records for the future. While other formats or systems may already capture the same information, this metadata should be crosswalked or added to PREMIS for archival storage to ensure future interoperability.

The PTF assumed that any PREMIS preservation metadata to be recorded pertains to master copies that are placed in archival storage but not to their access derivatives. While the PREMIS records for the master files can and should contain links to access derivatives (primarily through the use of <linkingIdentifier> elements), the derivatives themselves do not need to be tracked in additional PREMIS records.

In addition to storing basic technical metadata, the PREMIS record can record the occurrence of several preservation activities. The PTF agreed that PREMIS should track some basic preservation activities. Some of these activities, such as the initial ingest, routine fixity checks, and the refreshing of media, should include the capture of a checksum of the bitstream as well as the date and time of the event. Other events recommended for tracking included creation of a new derivative, deletion of a bitstream, or the creation of contextual relationships between bitstreams (such as when a transcription text is added to an aggregate object of images from a book).

These events were chosen based on the options available in the PREMIS events category and the abilities of several tools that were also studied, not on activities that were being performed routinely already. The PTF felt that they should be articulating not only what PREMIS could track, but also what *should* be tracked even if that would mean developing new activities in addition to implementing the standard.

Technical Requirements and Metadata Tools for Implementation

In order to implement PREMIS and track the events listed above in a meaningful way, the PTF defined technical requirements and investigated tools for use in PREMIS creation. The PTF recognized that, while not presently available to library staff, the following mechanisms for PREMIS creation and management would be required: a tool for collection owners to create PREMIS records as well as an exchange mechanism or uploader for submission of PREMIS records created by another tool.

In addition to describing requirements that the library would need to build and implement, the PTF also investigated the following tools known to produce PREMIS records.

DAITSS

DAITSS Description Service, a tool developed by the Florida Center for Library Automation, uses DROID and JHOVE to characterize files and then ports this information into a PREMIS record.[3] The PTF found the PREMIS record generated is not as robust as needed for reformatted archival content, and attendant event and agent information would need refining. Use of this tool in AV workflows would also be restricted to audio files, with possible use for batch analysis and initial generation of PREMIS records.

MediaInfo

MediaInfo is a powerful tool for technical metadata extraction from video and audio files.[4] Mac, Windows, and command line versions of this tool exist, each with a different level of functionality. However, the technical information it extracts is quite thorough and fits the concerns of the archivist. Batch processing of files is possible in the command line version, but requires minor additional scripting. Output options for MediaInfo's technical report include XML or CSV options (only in Windows and command line versions). The potential for batch analysis would make it ideal for readdressing existing digital content for which there is inadequate technical metadata.

PBCore Instantiationizer

The PBCore Instantiationizer, a tool built by an audiovisual archivist, is a simple GUI utilizing MediaInfo's libraries.[5] It uses an XSL document to translate MediaInfo's XML output into PBCore elements, creating a PBCore instantiation record for the analyzed file. The technical metadata generated using this XSL, formatted as PBCore instantiation information, could be pasted directly into PREMIS' objectCharacteristicsExtension. The PTF determined that use of this tool would likely be restricted to video content, for which PBCore is an attractive option for both technical and descriptive metadata.

Archivematica

While not a tool created specifically for the creation of PREMIS, Archivematica creates a PREMIS record as part of the package of material required for storage in an OAIS-based repository.[6] As such, Archivematica accepts SIPs (submission information packages), or packages of original content to be submitted. Then it performs several functions on the files to prepare them for long-term storage as an AIP (archival information package). It also prepares a DIP (dissemination information package), which is a copy of the same files, in some cases migrated to preferred access formats. An overall METS record is created for each AIP which includes a complete PREMIS section for each file. This PREMIS record includes object characteristics, agents, and tracks the events carried out by Archivematica.

While none of the tools examined were perfect, they each created basic PREMIS records that included object characteristics. However, but the PTF was now envisioning the creation of PREMIS to be an ongoing activity, not a discrete step. PREMIS support would ideally need to be built into repository and archival storage practices and that would make implementation a much larger project.

Recommendations for Level of Implementation

With the standard assessed, tools examined, and collections available for PREMIS implementation, the PTF defined a tiered approach to applying PREMIS metadata in digital preservation activities. A flexible standard, PREMIS proscribes few requirements for conformance to the semantic units and data model. There are a relatively small number of elements listed as mandatory in the specification, and the obligations, constraints, usage notes, and definitions prescribed in the PREMIS Data Dictionary are straightforward and easy to follow. To create more guidance for the local implementation of PREMIS, the PTF developed a set of implementation levels that correspond with established "preservation levels" set by the Digital Curation Services unit as part of a research project in 2004.[7] The corresponding levels of PREMIS implementation were outlined as follows:

First Degree

All PREMIS-mandated semantic units are included at this basic level. In addition, files and bitstreams should undergo fixity checks at dark-archive ingestion and have resulting information recorded in PREMIS.

Second Degree

This level of PREMIS implementation includes all of First Degree PREMIS, as well as enough details about relationships among objects, and links to other PREMIS events and agents, to orient a digital object within some larger structure or context. Noting the relationships among files, bitstreams, representations, and intellectual entities, as well as the linking identifiers among PREMIS objects, events, and agents, is critical to tracking the digital provenance of objects.

Third Degree

At this level, the PTF suggests recording information about a digital object's creation circumstances as well as its rendering requirements, in addition to all aforementioned data. Preserving specific information about the object, what created it, and what would be needed to re-create it helps to ensure a reasonable reproduction of original content. In addition, a few of the semantic units required in previous levels are supplemented with additional sub-units.

Fourth Degree

Populating all PREMIS entities and their semantic units offers the best support for any object. Fourth Degree PREMIS provides a digital artifact's best chance at survival and future usability—it entails the entire PREMIS schema, or as much of it beyond the Third Degree PREMIS level as one might wish.

Estimate the Impact of Implementation on Staff Workflows

In the final part its charge, the PTF assessed the impact of PREMIS implementation on the workflows of both content creators and programming staff.

Ideally, the impact of PREMIS creation during the acquisition or digitization period should be minimal. Given that some of the technical metadata that contributes to preservation metadata is already captured as part of routine metadata collation for some departments, extra levels of metadata capture will not burden those content creators. However, in many other cases new tools will need to be identified, created, or modified for the capture of this data, and the PTF recognized that programming support would be needed to ensure that the output of these processes could be integrated with preservation repository services.

The PTF recommended that the Library should also commit programming staff to the development of digital preservation maintenance activities—at minimum, the systems must be able to take XML from various tools, track preservation activities, and track significant changes to access copies. There will be a large impact on programming staff if the archival systems, as they exist presently, are not able to meet these requirements. Programming staff will therefore need to work to create new tools at the repository level to accept and manage PREMIS metadata.

Conclusion

The PTF recommended PREMIS adoption to library management staff for a number of reasons, not the least of which was the national recognition of the standard and its demonstrated application in numerous projects at various institutions. In addition to the community of practice surrounding the standard, the PTF recognized the utility of the standard in capturing preservation metadata elements essential to a robust digital preservation plan.

Since submitting the report to the task force conveners, the library has recognized the need for the establishment of a sound digital preservation protocol as a precursor to any PREMIS implementation. This has, perhaps, been the most important result of the PTF's investigation. Digital preservation stakeholders at the library have come to understand that implementing preservation metadata for digital collections is not merely a matter of applying an additional metadata standard to content. Rather, digital preservation requires a robust and well-considered program of activity. A program for digital preservation must include repository management, the collection curation, proper tools, and trained staff that can carry out preservation activities in addition to preservation metadata.

The PTF and others with digital preservation concerns at the library agree: PREMIS is a good and viable option for use in a digital preservation program with substantial institutional, technical, and staff support. The act of engaging in evaluating PREMIS will lead to the development of better preservation services.

Notes

1. Priscilla Caplan, "Understanding PREMIS," Library of Congress Network Development and MARC Standards Office, (2009), http://www.loc.gov/standards/premis/understanding-premis.pdf, (accessed 14 February 2013), 3.
2. Priscilla Caplan, "Understanding PREMIS," 8.
3. DAITSS Digital Preservation Repository Software, Florida Center for Library Automation, http://daitss.fcla.edu/, (accessed 14 February 2013).
4. "MediaInfo," http://mediainfo.sourceforge.net/en, (accessed 14 February 2013).
5. "PBCore Instantiationizer," http://www.avpreserve.com/pbcore-instantiationizer/, (accessed 14 February 2013).
6. "What is Archivematica?" Archivematica, https://www.archivematica.org/wiki/Main_Page, (accessed 14 February 2013).
7. Sustaining Digital Scholarship, "Levels of Collecting," http://www.digitalcurationservices.org/sustaining-digital-scholarship/ (accessed 14 February 2013).