

Summary Report of the PREMIS Implementation Fair

The PREMIS Implementation Fair, sponsored by the Library of Congress, was held on October 7, 2009 in San Francisco. There were over 40 attendees and registrants from 14 different countries. The day was divided into eight sessions focused on the status of PREMIS, implementation in METS, tools, systems, changes in the data model, two sets of case studies, and an open discussion of implementation issues.

1. Status of PREMIS (Brian Lavoie, OCLC Research)

Brian Lavoie reviewed the progress of PREMIS from a conceptual framework in 2002 to the initial publication and revision of the PREMIS Data Dictionary in 2005 and 2008, respectively. Since completing the revision of the Data Dictionary, the activities of the PREMIS Editorial Committee has conducted tutorials, fostered tools, articles and documentation, developed PREMIS-METS guidelines, and is now addressing PREMIS conformance. PREMIS is the de facto standard for preservation metadata. Implementations of PREMIS are appearing in many forms but there is little consolidation of experience and best practices. Directions of future PREMIS work include: documenting and sharing experiences using PREMIS; working on new and existing tools to automate generation of standard preservation metadata; developing registries of shared data; harmonizing PREMIS with other standards (e.g. METS); and deciding who should collect preservation metadata and when it should be collected.

2. Implementation in METS (Rebecca Guenther, Library of Congress)

Rebecca Guenther highlighted the complex decisions involved in implementing PREMIS in METS. Because of the flexibility of both schemas and the need for greater interoperability, it was clear that guidelines for using PREMIS in METS would be useful. A group of PREMIS and METS experts was convened and the result of the group's work, the *Guidelines for Using PREMIS in METS for Exchange* is now available in draft form on the PREMIS website. The *Guidelines* address areas including: where to record PREMIS information in METS; managing PREMIS-METS redundancies; structural information; ID/IDREF mechanisms; PREMIS extensibility mechanisms for format-specific metadata; and documenting decisions in a METS profile. There are some potential future changes to the *Guidelines* and related changes to the PREMIS and METS schemas. The PREMIS Editorial Committee hopes to change the schema by the end of 2009 to better sync PREMIS and METS extensibility mechanisms. The Editorial Committee may also examine allowing URIs or strings to identify metadata. In terms of the *Guidelines* themselves, there are plans to add to use cases, examples, and more structural information.

3. Tools

PREMIS in METS Toolkit (Francesco Lazzarino, Florida Center for Library Automation)

Francesco Lazzarino described and demonstrated the three tools in the PREMIS in METS Toolkit recently developed by FCLA for the Library of Congress. The first tool, based on DROID and JHOVE, takes a file and describes it in PREMIS 2.0. The second tool takes a METS document and converts it into a PREMIS document or vice versa. This tool is implemented in XSLT and the code is available for open use. The third tool validates a document according to the PREMIS and METS schemas and the *Guidelines for using PREMIS in METS for Exchange*. This tool was developed using Schematron and that code is also available. PREMIS-in-METS Toolkit is a nearly stateless, RESTful web service. Though its interface does not support batch processing, the output of the web service and tools is very clean, making it possible to use PERL or other tools to implement batch processing.

Hub and Spoke Framework Tool Suite (Bill Ingram, University of Illinois Urbana-Champaign)

The NDIIPP-funded Hub and Spoke project has developed a single object model that can serve as a translational tool between different repository object models, making custom transforms unnecessary. In its profile, Hub and Spoke mandates use of PREMIS 1.0. There are no current plans to implement PREMIS 2.0 as the grant is over but UIUC may build a preservation repository that uses the Hub and Spoke tools which would implement PREMIS 2.0. Ingram gave a functional overview of the Tool Suite, which supports DSpace, Eprints, and Fedora and is built to be extensible. Major components of the Tool Suite's functional workflow include: an ingest web service; a packager that creates the Hub and Spoke

package from another repository's DIP or from content files and metadata; scripts that create technical metadata and PREMIS information; and a SIP creator. Ingram described the technical infrastructure of the Tool Suite and noted that it is open, documented, and that the METS profile is registered with Library of Congress and the DLF Aquifer profile of MODS is used.

Statistics New Zealand PREMIS Prototype Tool (Euan Cochrane, Archives New Zealand)

Statistics New Zealand identified a need for a low-cost tool that could combine the functionality of the NLNZ metadata harvester, JHOVE, and DROID to create PREMIS object records. Cochrane developed a tool using DROID XML output, JHOVE XML output, a filter script, NLNZ XML output, an XSLT script, and a splitter script to create individual PREMIS 1.0 object records. There are no plans to implement PREMIS 2.0 but Statistics New Zealand is waiting for funding to build a more robust tool which would implement PREMIS 2.0. A potential problem with the PREMIS prototype tool is that it may not be scalable as it is all XML and text based. The new FITS (File Information Tool Set) developed at Harvard University Library is much “bigger” but it does not use PREMIS or XML. Cochrane emphasized that the PREMIS creation tool is useful, that tools like it are easy to make, and that the tool shows the power and simplicity of XML.

4. Systems

PREMIS Use in Rosetta (Yair Brama, Ex Libris)

Rosetta is a digital preservation system developed in conjunction with the National Library of New Zealand. Version 1.0 was released in January 2009 and Version 2.0 will be released at the end of 2009. The Rosetta data model is based on PREMIS and the system is compliant with PREMIS and METS. In Rosetta, each METS document describes an intellectual entity consisting of several representations. Technical metadata in Rosetta is recorded using DNX, Ex Libris' own schema which includes and normalizes format-specific technical metadata from multiple schemas including MIX, textMD, and it can include PREMIS information. There are some redundancies between DNX, METS, and PREMIS. While Rosetta does not yet have the ability to convert from a DNX file to a PREMIS record, it wants to create that functionality as it Ex Libris sees that a solid PREMIS community has coalesced.

After Brama's presentation there was a discussion of schema versioning. There is no versioning strategy for Rosetta's DNX files but they will now consider that. Evan Owens of Portico noted that Portico is already facing problems with versioning XML and in response, has dropped use of enumerated mdTypes and includes versions of metadata types. Owens predicted that this issue will soon be of wide concern.

DAITSS (Priscilla Caplan, Florida Center for Library Automation)

The Florida Center for Library Automation (FCLA) has developed DAITSS, a dark archive software application designed to implement normalization and forward migration at time of ingest and only at ingest. DAITSS 1.0 is partly PREMIS conformant and DAITSS 2.0 will be entirely PREMIS conformant. In DAITSS, AIPs and DIPs implement PREMIS in METS. If a representation is normalized or migrated on ingest it becomes a new representation. Recording PREMIS events is crucial to ingest processes and external events (e.g. from a TIPR package) can be ingested with SIPs. Because of a design constraint necessarily tied with one of DAITSS' foundational principles (nothing but ingest can touch the data store), DAITSS records post-ingest events but does not add them to the object until after a dissemination.

In discussion, Caplan noted that currently, DAITSS is just disseminating on request, but that they can easily do mass migrations by doing mass disseminations. Caplan also noted that DAITSS obtains schema needed for validation recursively. PREMIS and METS schemas, however, are stored in a global files registry rather than in each AIP. Evan Owens noted that in Portico a separate section of their archive holds the DTDs, etc. and when Portico does a dissemination, they grab the schema out of that separate stash. Rebecca Guenther asked what happens when the PREMIS Editorial Committee or METS Board tweaks the schema but does not change the schema version. Owens responded that Portico checks for

changes and assigns their own version numbers to distinguish between tweaked versions of the schema because this causes major problems.

Rob Sharpe of Tessella remarked that their Safety Deposit Box system allows ingest of half a million files per day and that they do not currently use PREMIS or METS because it increases the size of the XML multifold and slows processing. He also noted that Tessella did something in some ways similar to DAITSS before the development of PREMIS. However, it is not PREMIS compliant and they wonder how they can leverage what their information, which includes additional non-PREMIS significant properties and object relationship information, in order to attain conformance. Caplan and Lavoie noted that if Tessella has preservation metadata that they find useful and is not in PREMIS, the Editorial Committee needs that feedback and that the PREMIS Implementors' Group (PIG) list would be a good venue for it.

When Caplan noted that DAITSS does not record environments, a participant noted that the Unified Digital Formats Registry (UDFR) expects to export PREMIS conformant environments information soon. It was observed that the environments issue has a lot to do with conformance and Rebecca Guenther noted that the UDFR is incorporating environments as a result of their interaction with the PREMIS conformance group.

5. Changes in data model

Intellectual entities; significant properties (Priscilla Caplan, FCLA)

Priscilla Caplan presented some suggestions the PREMIS Editorial Committee has received for changes to the PREMIS data model asked for comments.

The first proposed change concerns intellectual entities and has arisen from the Toward Interoperable Preservation Repositories (TIPR) project. In the TIPR model, DIPs are transformed into a neutral exchange format, RXP. Each RXP contains both package level and representation level information. TIPR plans to include 2 layers of PREMIS documents: one at the package level and one at the representation level. The package level is roughly equivalent to the Intellectual Entity, and PREMIS object description at this level is not supported by the current PREMIS data model. Therefore, TIPR is asking the PREMIS Editorial Committee to consider expanding PREMIS applicability to include metadata for higher level objects like packages.

The second proposal concerns significant properties and stems from the work the British Library has done. The British Library has come up with a more nuanced view of and data model for significant properties, which they have termed "significant characteristics." They have distinguished three different levels of objects: physical object level (bytestream); representation level (files and representations); logical level (intellectual entity and components of objects that share preservation characteristics). An example of the logical level is a JPEG image that could be a standalone file or embedded in an article. Whether the JPEG is alone or embedded, it shares preservation characteristics with other JPEGs. Because components of objects may share characteristics, the British Library is examining applying significant properties not to files and bitstreams, but to logical components, which implies that components should be entities in the PREMIS model.

Portico: ITHAKA Preservation Metadata 2.0: Revising the Event Model (Evan Owens, Portico)

Evan Owens framed his presentation not as a proposal for the PREMIS Editorial Committee but as a use case for thinking about events. The Portico archive is very large (150 million files, 1 billion events) and its design was heavily influenced by PREMIS. In a recent metadata review, Portico decided to use their own data model, largely to reduce versioning problems by using XSD schema. They use their own schema, not PREMIS or METS, but they are PREMIS compliant. To reduce redundancy, they decided to create a master record for batch processes and link individual objects to that master record rather than recording details of event at the object level. Portico is currently working on a controlled vocabulary of event types. Owens observed that large scale events feel very different from human events and due to

Portico's scale, they consider every bit of metadata. However, recording events has proven its value and has been invaluable in recovering from damaging mistakes. In response to a question about how to make the case for events tracking to administrators, Owens noted that one of his mottos is: Hardware is not perfect, software is not perfect, people are not perfect and a backup strategy and what you have recorded about what you have done are your only protections against loss.

6. Case studies

Using PREMIS to automate rights management (Bradley Westbrook, University of California San Diego)

Management of digital object rights at UCSD posed an access problem. Because rights were managed at the collections and were dealt with in an ad-hoc manner, material that actually should have been available was restricted and vice versa. There was no procedure for managing restriction expirations. Consequently, the Digital Asset Management group implemented object-based rights metadata and switched from using METS rights to PREMIS rights because it was more expressive. In the new rights workflow, every object has to have copyright information and licensing information when applicable. A few processes use this metadata. An access status determination process divides the collection into objects that can and cannot be shown and a check ensures that the rules are properly applied. Access status can be overridden when, for example, they know a particular collection is out of copyright, even if copyright status is officially unknown. In the new system, rights metadata is automatically updated and does not rely on manual metadata revision and additions. The access status process is executed roughly every 30 days to re-determine item availability.

In response to questions, Westbrook noted that changes to rights are not being tracked as PREMIS events; that they expect to provide additional copyright status options, such as Creative Commons, if useful; and that while display is currently binary and set to either public display or no display, they expect to support a range of other options in the future including, read, write, and edit.

Implementation in Italy (Angela Di Iorio, Fondazione Rinascimento Digitale)

Di Iorio works with a small private foundation that supports digital preservation projects in Italy. The Fondazione has started a project to exchange AIPs. This project, the ARTAT (Archives Ready to Transmit AIPs) has three partners: the Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information, which holds the institutional repository MAGTECA for the Italian National Digital Library Portal and Cultural-Tourist Network (use the Italian metadata standard MAG); Magazzini Digitali, a Fondazione Rinascimento Digitale and National Library of Florence project to archive doctoral theses (use MPEG21-DIDL); and the digital repository of Library and Archive of the British School at Rome (use METS). In first, inquiry phase of the ARTAT project they collected information from repositories. The ARTAT project then developed the concept of a Preservation Metadata Layer and requirements for the structure of the Preservation Metadata Layer (PML). The PML is independent of AIPs and at its core, is data about the metadata of an AIP. The PML require PREMIS conformance and comprehensiveness. ARTAT has established a semantic unit roadmap.

In response to questions, Di Iorio said that they are starting with these three partners in 2009 and are looking for future partners and would be happy to see this model used in other countries if there is interest.

PREMIS & Geospatial Resources (Nancy Hoebelheinrich, Knowledge Motifs)

Hoebelheinrich addressed PREMIS in the domain of geospatial data. She reviewed the characteristics of two types of geospatial data, high resolution orthoimagery (HRO), which uses shape files, and compiled GIS datasets, which compile different data points from various layers of data. The complexity of geospatial data necessitates answering tricky question about what to archive, the purpose of the data, the context, and how much of context to keep. For PREMIS, geospatial data presents structural, contextual, provenance, and events complexities. NASA and NOAA are pushing a lot of change in this domain and PREMIS may eventually be built into geospatial tools. Geospatial data may require some exploration of

PREMIS agents and significant properties. Groups and projects to watch in this area include the ESIP Federation, which is developing a metadata testbed, the American Geophysical Union, and the NDIIPP geospatial projects.

Discussion started in response to a question about granularity and levels of data and at what point you would want to archive certain files (e.g. layers? shapefiles?). Another participant commented that this is same problem in the changes to the data model discussion, the TIPR problem of keeping metadata at a higher level than PREMIS currently supports. In the geospatial cases, there is also the familiar issue of files that get re-used and needing to decide if they will be defined as different objects or versions of an object.

7. Case studies II

PREMIS in TIPR (Priscilla Caplan, FCLA)

The TIPR project is an IMLS- funded partnership between FCLA, Cornell and NYU, who all use different systems (DAITSS, aDORe, DSpace). TIPR's goals are to demonstrate the feasibility of repository to repository transfer of enriched AIPs and to identify issues that impede transfers. In TIPR, repositories must maintain digital provenance through transfer and must understand and be able to use PREMIS and METS information from another repository. TIPR mandated PREMIS and METS because they wanted to use existing standards and because they wanted the project to be as "lightweight" as possible. PREMIS also proved to be useful for discussing differences in AIP structures. TIPR has made significant progress thus far. They have defined a package structure, RXP, for exchanging information. The RXP includes both package-level and representation-level PREMIS information. Each partner can take a DIP from their repository and turn it into an RXP as well as ingest an RXP from other partners. Future work for TIPR includes: discussion with the PREMIS Editorial Committee about the support for higher (e.g. package) level information in the Data Dictionary and a "round-trip" transfer, in which one repository transfers a package to another repository, which then ingests it and disseminates and transfers it back to the source repository.

During discussion, Caplan noted that TIPR felt that for an exchange, the receiving repository did not have to understand the package's descriptive metadata. However, partners can agree to do and mutually understand more than is specified in the RXP. The TIPR group designed the RXP so that cases like disaster recovery would not be excluded due to overly-prescriptive rules. TIPR and the Hub and Spoke Tool Suite offer an interesting contrast in approaches. Hub and Spoke has the same goal as TIPR, but assumptions about what the receiving repository needs to know are quite different.

National Library of Finland digitized monographs (Karo Salminen, Jukka Kervinen, National Library of Finland)

As part of a National Digital Library Initiative, the National Library of Finland (NLF) recently decided to revise their digitization workflow and extend their METS profile to include PREMIS. The NLF wanted their METS profile to capture digitization agents and events. Each NLF METS package includes a PDF, and for each digitized page, a master, access, and thumbnail image with OCR text in ALTO XML. Currently this profile is not an implementation but a "wish list." Salminen outlined some details of the NLF METS profile, which spreads PREMIS information across different METS amdSecs. The NLF distinguishes between file-specific events and events that are common to all files in an intellectual entity. Some currently unresolved issues include redundant information in PREMIS, METS, and MIX, and the size of the metadata (a monograph of 300 pages, would have roughly 400,000 lines of XML in file).

PREMIS Implementations at the British Library (Markus Enders, British Library)

The British Library (BL) thinks of AIPs as conceptual entities that are part of a write-once system, i.e. any change means updating the whole AIP and then submitting it as a new AIP. PREMIS is being used for three different content streams at the British Library. The first stream, e-journals, implements PREMIS 1.0. There is one AIP per article, issue, journal, and digital manifestation. For issue, journal, article representations, AIPs only consist of a METS descriptor. For digital manifestations, AIPs consists

of content files and a METS descriptor. The next two content streams, newspapers and web archiving both use PREMIS 2.0 and required changes to the AIP structure. A hierarchy of AIPs was no longer feasible and instead, there is one AIP per newspaper issue and there could be several manifestations per AIP.

The British Library have encountered problems with controlled vocabularies; non-validatable semantic dependencies between METS and PREMIS; difficulty of creating a unified workflow for creating PREMIS and METS documents rather than content and structure-specific workflows; and recording preservation metadata on metadata.

SCHEDULE CHANGE: Because the case studies ran over on time, discussions of PREMIS conformance (Brian Lavoie) and controlled vocabularies (Rebecca Guenther) were eliminated from the agenda, but the slides are posted online at the PREMIS Implementation Fair site.

8. Open discussion of implementation issues (led by Rebecca Guenther)

Nancy Hoebelheinrich asked if anyone was testing to see how helpful the preservation metadata they are gathering actually is in reconstructing objects. Priscilla Caplan responded that most implementors probably did not test the value of their metadata and that such a measurement project would make a great grant opportunity. An attendee mentioned Steve Abrams' study of JPEG conversion, in which he concluded that his program was not recording the right significant properties. Evan Owens said that Portico is a good test of the business case for preservation metadata. However, the business case has expanded beyond future replication of data as PREMIS has proven useful for managing repositories. Owens observed that there seems to be a worrisome assumption that no two PREMIS implementations are identical, which gets back to flexibility and archive management problems.

The discussion also covered environments. Guenther asked if environments should be: part of objectCharacterites, part of events, or a separate entity. A participant questioned the necessity of storing environments and what kinds of environments we would store: rendering, creation, or ingest environment. Depending on the environment, you may want to generate it on the fly or point to a registry such as UDFR. One argument for splitting environment information is the need, when the data allows, to push metadata to higher levels of objects to take advantage of classes.

There was also discussion of deciding how to store or reference files that are frequently used in conjunction with AIPs, for example the METS or PREMIS schemas. Markus Enders noted that this file redundancy problem is not a problem of the data model, but a question of serialization. There is a difference between the Data Dictionary, which is more flexible, and the schema. One participant observed that perhaps we need a "BagIt" type mechanism that references objects referenced many times in METS or PREMIS.

Evan Owens noted that in 5 years the PREMIS Implementation Fair attendees will be dealing with problems well beyond ingest and will likely have whole new lists of events, new problems, and new case studies to bring to such an event. Owens predicted that the "killer issue" of the next 5 years will be exchange. We can agree on event names, on using MODS, but coming to standard structural assembly rules is going to be much trickier.