Implementing PREMIS in Container Formats

Rebecca Guenther, Library of Congress, Washington, DC
Zhiwu Xie, University of New Mexico and Los Alamos National Laboratory, Los Alamos, NM

Abstract

In May 2005, the PREMIS Working Group released the first version of the Data Dictionary for Preservation Metadata: Final Report, a community consensus based preservation metadata standard. PREMIS XML schemas were also published to support the implementation. Since then, many organizations started implementing PREMIS in their repositories, during which a handful of common implementation issues surfaced. The community is finding a variety of solutions to these issues and anticipates the emergence of best practices. The PREMIS Implementers' Group (PIG) was formed after the release of the data dictionary to address implementation issues through open discussions and consensus building processes. In August 2006, the PREMIS Editorial Committee was established to coordinate and approve future revisions of the Data Dictionary and XML schema, as well as provide guidance for implementation. This paper summarizes some of the experiences gained in implementing PREMIS in container formats, especially in the context of the METS and MPEG-21 DID frameworks. Attempts are made to draw best practices from our experience for future implementers.

Introduction

Born digital and digitized information has formed an ever greater portion of human knowledge. Although providing many unsurpassed advantages over paper media, digital assets are indeed more vulnerable to becoming obsolete under rapid technological and managerial changes. The urgency of digital preservation has been widely recognized, and a handful of national and international efforts have attempted to address various aspects of the issue. PREMIS (PREservation Metadata: Implementation Strategies) is the first international effort to standardize the preservation metadata, the information that supports and documents the long-term preservation of digital materials.

Convened by OCLC and RLG, the PREMIS Working Group developed the PREMIS Data Dictionary [1], which contains a comprehensive view of information needed to support digital preservation activities, including recommendations for guidelines and recommendations to support creation, use and management of digital resources. It is based on a deep pool of institutional experiences in setting up and managing operational capacity for digital preservation. The PREMIS Working Group received two awards for its work, the UK 2005 Digital Preservation Awards and the 2006 Society of American Archivists' Preservation Publication Award, illustrating the attention it has received from a broad audience.

By definition, the preservation metadata addresses an archived digital object's provenance, authenticity, preservation activity, technical environment, and rights management. The PREMIS standard 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. It is part of the necessary infrastructure for a sustainable digital preservation activity.

There were several guiding principles in the development of the PREMIS Data Dictionary. In order for the group to make progress in a reasonable period of time, some limitations on the scope were necessary. For example, the Data Dictionary only includes technical metadata that are applicable to all types of file formats, rather than format specific technical metadata. Business rules of a repository were deemed out of scope, although they play a key role in managing digital objects. In addition, there was an emphasis on automated workflows, which was necessary to define an implementable set of metadata, since resources would not allow for much human interaction given the increasing number of digital objects.

The PREMIS Data Dictionary was intentionally technically neutral in that no assumptions were made about specific archiving technology, system or database architectures, or specific preservation strategies. In terms of metadata management, no assumptions were made about whether the data was stored locally or externally, how metadata units were instantiated, or whether values were recorded explicitly or known implicitly. The term "semantic unit" was used instead of "metadata element", further emphasizing this technical neutrality, since "metadata element" implies implementation in a structure like XML. It was felt that this approach would promote flexibility and the ability to apply the specification in a variety of contexts. However, with this flexibility comes the need for further guidance and the development of application specific best practices.

The PREMIS Data Dictionary was released along with five XML schemas to allow for its implementation. The schemas were faithful to the semantic units specified in the Data Dictionary in terms of characteristics such as naming and repeatability. They were defined in terms of the entities in the PREMIS data model, which include Object, Event, Agent, and Rights. Semantic units in the Data Dictionary are organized in terms of these entities. There is one XML schema for each entity and a PREMIS container schema is also provided to wrap all four related schemas if desired.

Since the release of the data dictionary, a handful of organizations have started experimenting, evaluating, and implementing PREMIS. Many others have shown interest or are actively working towards adopting the standard. A series of tutorials have been held in various locations, and more are planned. The PREMIS Implementers' Group was formed after the release of the Data Dictionary to address implementation issues through open discussions and consensus building processes. In August 2006, the PREMIS Editorial Committee was established to coordinate and approve future revisions of the Data Dictionary and XML schema, as well as provide guidance for implementation.

This paper documents two exemplary PREMIS implementations, in the form of PREMIS elements within two different XML containers, namely Metadata Encoding and Transmission Standard (METS) [2] and MPEG-21 Digital Item

Declaration (DID) [3]. It is expected that using PREMIS within XML container formats could satisfy the need for standardized exchange formats. We summarize the experiences gained during the development of these examples, and attempt to draw best practices for future implementers.

OAIS Reference Model and Containers

The Open Archival Information System (OAIS) reference model [4] was developed by the Consultative Committee for Space Data Systems (CCSDS) to provide a conceptual framework and common vocabulary for digital preservation activities. The model was approved as ISO standard 14721 in 2003 and has gained wide recognition among the community. Besides the functional model for preservation activities, the OAIS also includes an information model specifying types of information required for long-term preservation. These information objects are then conceptually grouped together to form an information package.

The development of the PREMIS standard used the OAIS reference model and more specifically, its information model, as a starting point. The PREMIS Data Dictionary consolidated and further developed the conceptual types of information objects into more than 100 structured, logically integrated, and implementable semantic units, and more importantly, provided detailed descriptions and guidelines to implement them.

While the PREMIS metadata is detailed and thorough for the preservation purpose, it is important to recognize they are only a subset of the metadata associated with an intellectual entity. The real-world implementation of an Archival Information Package (AIP) may include much more metadata besides the preservation metadata, therefore a well defined container is usually necessary to group and appropriately associate these metadata with the data object. In this paper we focus on the XML implementation of the PREMIS, therefore the container formats are also in XML.

METS

The Metadata Encoding and Transmission Standard (METS) [2] is an XML schema that provides a standard encoding for descriptive, administrative, and structural metadata for objects in a digital library. It records the hierarchical structure of digital objects, the names and locations of the files that comprise those objects, and their associated metadata. As such, the METS schema serves as a container to contain both metadata and files or links to files. A METS document may be a unit of storage or a transmission format, in the sense of an OAIS submission information package, archival information package, or dissemination information package.

METS is extensible and modular by providing places where elements from other XML schemas can be plugged in. These are called "extension schemas", and the METS Editorial Board endorses some, while others may simply use the XML schema facility for combining vocabularies from different namespaces. The METS schema itself only defines what is contained in the METS header, the file section (with names and locations of files that are part of the METS document), structural links (internal document), the structural map (laying out the structural relationships between files and parts of the METS document), and the behavior sections (associating executable behaviors with content in the METS object), while the descriptive and administrative metadata sections bring in elements from other schemas. Subsections of the administrative metadata section include technical metadata (detailing the file's creation, format and use characteristics), digital provenance metadata (detailing source/destination relationships between files and actions performed upon objects detailed within the METS document), and rights metadata (expressing copyright and license information).

MPEG-21 DID

Although less known to the digital library world, the Moving Picture Experts Group (MPEG)'s MPEG-21 DID standard [3] has become a competitive alternative as a digital object container [5]. It has also been mapped to the OAIS information model and shown to be a suitable candidate for the AIP implementation [6].

Designed to realize the universal multimedia access, this new generation of ISO standard suite carefully crafted its data model to be extremely flexible and interoperable with external standards. Within this data model, the descriptor/statement structure may be used as a metadata container. This structure can be attached to the DID container itself (therefore mapped to the package information), or to an item and its component(s) as other metadata. The items and components can be infinitely nested, therefore well accommodating the requirements posed by compound objects with complex inner structures. The descriptor/statement structure is designed to accept well formed XML from any namespace, therefore allows metadata from external namespaces to be appropriately organized around the data objects at various levels of the information hierarchy.

In comparison to MPEG-21 DID, METS metadata appears to be more explicitly segmented. The categorization of metadata into different sections of descriptive, administrative, structural metadata is helpful in organizing the types of metadata required, but can be misleading when the distinctions between desriptive, structural and various forms of administrative metadata are blurred. [7].

Implementing PREMIS in METS

Because METS defines subsections under the administrative metadata section, a choice must be made as to where to include the preservation-related metadata from the PREMIS schemas. As noted above, there is an XML schema for each entity in the PREMIS data model as well as a container schema, which references each separate schema. In general, the Object entity contains technical metadata about the object, so would be appropriately used in the METS techMD section. The Event entity contains metadata about actions performed on the objects, so it would be appropriate in the METS digitalProvMD section (i.e. digital provenance). Likewise, the Rights entity contains rights and terms and conditions metadata, so could be used in the METS rightsMD section. The Agent entity contains information about an agent in terms of events or rights, so would be associated with either of those two entities.

On the other hand, if we consider all PREMIS entities as integral components of the preservation metadata which can also be considered as part of the digital provenance metadata, it would also be acceptable to put them all in the <digiprovMD> section. To do so, theoretically it may be preferable to first wrap these entities in the top-level <premis> container. Practices have varied so far, and in the actual repositories the <premis> wrapper may have not been implemented.

Best practices are needed for implementing PREMIS in METS because of the metadata categorization and the non-prescriptive nature of the METS schema. The implementer may make choices in terms of a number of questions, such as:

- Whether to use separate METS administrative metadata sections for each PREMIS entity and which ones subsections to use (i.e. techMD, digiProvMD, rightsMD), or to include all PREMIS metadata together in one subsection under the administrative metadata section (<adminMD>)
- Whether to use the <premis> container schema to wrap the metadata from separate XML schema(s)
- Whether to repeat any metadata which is defined in PREMIS and also defined as elements in METS (e.g. fixity in PREMIS; checksum in METS which are associated with files)
- How to record technical metadata elements which are defined both in PREMIS and in a format-specific technical metadata schema (e.g. Metadata for Images in XML schema (MIX)[8] for digital image technical metadata)
- How to record structural metadata, which in METS is carried in the <structMap> section and in PREMIS is contained in discrete elements in the Object schema under <relationship>. Since the <structMap> is required in METS, the question remains whether to redundantly record structural relationship information in PREMIS metadata as well.

METS implementers are considering these questions, and they will establish best practices based on implementation experience. It may be possible to use xPath or xPointer to associate metadata from different METS sections, although this would require an extension to the METS schema. There are a number of implementations that are exchanging METS documents, and this will facilitate such sharing of digital objects and their metadata. As of this writing, conclusive consensus has not been reached.

An example for the PREMIS implementation in METS can be found at:

http://www.loc.gov/standards/premis/louis.xml

This example puts each of the PREMIS entities in the METS section most appropriate to it, and do not use a PREMIS container, which would result in the repetition of some elements. As for the metadata repeated in both METS and PREMIS, the preference seems to be recording metadata redundantly defined in PREMIS and METS, since an application may wish to keep all PREMIS metadata together and not have to parse the METS document to find it.

Implementing PREMIS in MPEG-21 DID

Since MPEG-21 DID does not impose categorization on the metadata, where the top-level PREMIS container and the four entities should reside in DID is straightforward. Based on the OAIS mapping of the DID model [6], it is only natural that the overall PREMIS metadata, or the top-level <premis> container element itself, should be included in a descriptor/statement element at the uppermost <item> level of a DID file. On the other hand, since the different <object>, <event>, <agent>, and <rights> metadata can also be deemed as associated with different components, they may reside separately at the respective <component> level contained in the <item>.

Similar to the METS implementation, the question arises on how to deal with the metadata redundancy at the <item> and the <component> level. We can approach this issue from two different perspectives:

- Not limiting the redundancy issue within the context of implementing PREMIS in DID container, if a redundancy is due to schemas overlaps, that two or more different metadata schemas happen to contain elements of the same meaning and therefore the same value, then as long as each has its own rationale to exist independently from the others, we choose to treat them as necessary, and therefore leave them as is. One such example is the object identifier, which appears in multiple schemas. We do not attempt to reduce such redundancies.
- In some cases, the redundancies are from the same schema, not absolutely necessary, and can cause maintenance problems. For example, the same PREMIS <object> element may appear at both the <component> and the <item> level (within the PREMIS container). We address such problems from the perspective of best practice. Such repetitions can be reduced with either XLink or XInclude techniques.
- It is also important to notice that in many cases the repetition overhead is quite small, therefore cheaper to simply leave the metadata redundancies as is than having to update the XLink and XInclude whenever the XML structure changes.

An example for the PREMIS implementation in DID can be found at:

http://lakh.unm.edu/did.xml

This particular example has not included the <premis> container, although we recognize the necessity to implement it for a more complete PREMIS implementation.

We also went further to explore an unconventional, bottom-up approach to implement PREMIS. First we implement some lower level PREMIS semantic units without their PREMIS container. After we have accumulated enough lower-level PREMIS semantic units, we may gradually move up the ladder, until a full PREMIS implementation can be achieved.

The incentives for doing so arise from two practical implementation concerns. First, a full PREMIS implementation can be prohibitively expensive. Although only a few semantic units are required in the PREMIS data dictionary, it is always desirable to include as much optional metadata as one can automatically generate or collect. This requires careful planning, analysis, design, and in many cases considerable amount of implementation work. It also poses technological, financial, and managerial risks, therefore may deter potential implementers. A more practical approach should be available to the implementers so that a lower implementation threshold allows organizations to evaluate and experiment with the standard before fully adopt it. The PREMIS Editorial Committee is discussing the possibility of establishing a “PREMIS lite” implementation.

Second, many PREMIS metadata, such as those used to record software and hardware environment, carry useful information even without their original containers, and can be very useful for many purposes other than long-term preservation. The PREMIS data dictionary gives a well defined guideline on when, where, and how these metadata can be used, a feature many other metadata standards lack. When no suitable standard is available for that particular purpose, adopting PREMIS is indeed more attractive than inventing new ones.

This approach, however, requires the published PREMIS XML schema to be modified, because the existing official XML schema defines all lower-level elements locally, and therefore they are meaningful only within their direct containers. Based on the discussions on the PIG mailing list and a thorough examinination of the current schema, a new schema [9] was proposed and used as the basis for this partial MPEG-21 DID implementation.

The changes proposed in this new schema are currently under discussion within the PREMIS Editorial Committee. Many are expected to be adopted in the next official schema. Major changes proposed in this new schema are:

- Elements Globalization. All elements in the PREMIS schema are redefined as globally accessible instantiation of globally defined types. The semantics and syntax of the PREMIS Data Dictionary is strictly maintained, while including lower-level PREMIS elements in other containers becomes possible.
- Hierarchical <object> schema. The <object> schema is redefined in the way that each of the three categories of the objects is an extension to the abstract <object> element. Containing elements then are placed into the concrete types, each with their separate obligations and cardinalities.

After the modification, we can progressively implement PREMIS, not only in DID, but also in any other XML container schemas (e.g., METS) that allow inclusions of “foreign” elements. These changes also allow for more complete validation of usage by object category.

The given DID example illustrates this idea. The DID describes a digital object that consists of a pdf file as its component, and a nested item comprising of two components of xml files, one for the original descriptive metadata provided by the publisher, another for the MARC 21 record.

For the pdf file, the file format and size information are important and we do want them to be recorded somewhere. But we have not decided yet how far we want and can implement PREMIS at this time, so we simply implement the file format and size elements at the pdf component level, without implementing their containers.

Suppose we know more about the two xml components and decide to implement a more detailed PREMIS record on them, the example shows that the PREMIS <object> has been implemented for them.

Summary

This paper reports the experience gained during some early XML implementations of the PREMIS standard. Since the preservation metadata is only a subset of the metadata accumulated over time around the digital contents, it is usually preferred to wrap the PREMIS XML implementation within a container, which also houses other forms of metadata. The use of container is considered to be compatible with, and to some extent even a concrete implementation to the OAIS information model’s information package concept. This paper discusses PREMIS implementation issues in two of such containers: METS and MPEG-21 DID.

When implementing PREMIS in METS, variations exist on where to fit the four top-level PREMIS entities in the METS metadata categorization framework. Although the MPEG-21 DID implementation can circumvent this issue, how to make PREMIS efficiently and effectively co-exist with the other metadata remains to be resolved with best practices across both implementations. For now the consensus seems to be allowing repetition whenever justified.

This paper also discusses a bottom-up PREMIS implementation in the DID example, which requires the official schema to be modified. This approach provides a lowered implementation threshold, and allows PREMIS metadata to accumulate and mature gradually over time, first outside their PREMIS container, then put together to form a more complete PREMIS implementation.

References

[1] PREMIS Working Group, Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group (2005). Retrieved from http://www.oclc.org/research/projects/pmwg/premis-final.pdf

[2] The Library of Congress: The Network Development and MARC Standards Office, Metadata Encoding and Transmission Standard (METS) (2007). Retrieved from http://www.loc.gov/standards/mets/

[3] International Organization for Standardization, ISO/IEC 21000-2:2003. Information technology -- Multimedia framework (MPEG-21) -- Part 2: Digital Item Declaration (ISO, Geneva, Switzerland, 2003).

[4] International Organization for Standardization, ISO 14721:2003, Space data and information transfer systems -- Open archival information system -- Reference model (ISO, Geneva, Switzerland, 2003).

[5] Jeroen Bekaert, Emiel De Kooning and Herbert van de Sompel, Representing digital assets usingMPEG-21 Digital Item Declaration, International Journal on Digital Libraries, 6(2), (Springer: Berlin / Heidelberg, April, 2006)

[6] Jeroen Bekaert, Xiaoming Liu, Herbert Van de Sompel, Representing Digital Assets for Long-Term Preservation using MPEG-21 DID, PV 2005, Edinburgh, 21-23 November 2005.

[7] Brian Lavoie and Richard Gartner, Preservation Metadata, Digital Preservation Coalition Technology Watch Report No. 05-01 (September 2005). Retrieved from http://www.dpconline.org/docs/reports/dpctw05-01.pdf

[8] The Library of Congress: The Network Development and MARC Standards Office, Metadata for Images in XML Schema (MIX) (2007). Retrieved from http://www.loc.gov/standards/mix

[9] Zhiwu Xie, Jeroen Bekaert, and Herbert Van de Sompel, Proposed revision of PREMIS Schema (March 2006). Retrieved from http://purl.lanl.gov/aDORe/schemas/2006-04/

Author Biography

Rebecca Guenther has worked for the Library of Congress in various positions since 1980, and is currently a networking and standards specialist in the Network Development and MARC Standards Office. She works primarily on metadata, including the development and maintenance of MARC formats, as well as a number of XML formats, such as MARCXML, MODS, METS and PREMIS. Guenther is the chair of the PREMIS Editorial Committee and served as co-chair of the PREMIS Working Group during the development of the PREMIS Data Dictionary.

Zhiwu Xie is a member of the PREMIS Editorial Committee. He is currently a PhD candidate in the University of New Mexico, and works at the Los Alamos National Laboratory Research Library.