Ellen Phillips

Overview of the PREMIS data model

SJSU INFO 281

March 6, 2018

Introduction

The need for metadata schemas that can support long-term preservation of digital objects is clear. There are many to choose from, but the data model known as PREservation Metadata: Implementation Strategies (PREMIS) aims to suit the needs of most repositories. With it's relatively simple data model, PREMIS is easy to conform to. This is achieved by having relatively few entity classes containing only the most essential data needed for long-term preservation, while relying on other schema for description and structure.

The structure is based on semantic entities or units as opposed to metadata elements. This means that PREMIS does not specify how metadata should be formatted, rather it prescribes what types of information the system needs for preservation and exporting (Caplan, 2017, p.5). Some of the semantic units are merely containers to aggregate other multiple related semantic subunits (p.5). PREMIS also allows for Extension Containers. These do not have subunits, instead they are the designated spaces to record other types of metadata that are crucial to preservation but that are out of scope

of the PREMIS Dictionary (p.6). Additional structure is added through the use of outside vocabularies and implemented using Web Ontology Language (OWL).

Background

The first version of PREMIS was developed by the Online Computer Library Center (OCLC) and the Research Libraries Group (RLG) in 2005 and was later adopted by the Library of Congress (LOC). It consisted of the PREMIS Data Dictionary, a "comprehensive, practical resource for implementing preservation metadata in digital archiving systems" (OCLC & RLG, 2005 para.1). This release defined the data model, described limits and exclusions, and provided specific examples in XML. It has been widely adopted, and as Wilson (2010) observed "it has overshadowed other preservation metadata initiatives and seems to have swept all before it," (p.210).

PREMIS OWL

Release 2.2 included the adoption of the PREMIS OWL ontology. One of the things that made PREMIS easy to customize to local repository rules, also hampered interoperability. As Coppens et al. (2013) wrote, "there were a lot of free-text fields in the PREMIS formalisations, PREMIS OWL solves this by integrating 24 preservation vocabularies of the LOC," (p. 88). This

also allowed the data model to support more descriptive metadata in order to adhere to best practices like those as defined by the Open Archival Information System reference model (OAIS), (p.89).

With an included ontology, both preservation and dissemination needs are better supported because information packages can be fully described. As Coppens et al. explained, "using the Web Ontology Model (OWL) allows us to relate the entities to each other in a more harmonious way, because RDF is resource based," (p.90). The PREMIS Editorial Group (2013) stated that since OWL allows the data dictionary to become serialized in PREMIS it "can be leveraged to have a Linked Data-friendly data management function for a preservation repository." (para.3)

PREMIS OWL 2.2 is currently being revised. Major changes include a shift to include better practices around linked data. According to the PREMIS Editorial Committee, "this revision asserts relationships between PREMIS classes and properties and other vocabularies, and in some cases reuses external classes and properties," (2017a, p.3). The public may comment on the revised ontology until March 23, 2018 (PREMIS 3 OWL Ontology Draft Release, 2017b, para.4).

PREMIS 3.0

A major revision was released in 2015 that reclassed one of the main Semantic Units. The original model had five entities: Intellectual, Objects,

Events, Agents, and Rights. According to PREMIS (2015), in the first two versions, "intellectual entities were out of scope and semantic units to describe them were not included" (para.1) except to link to descriptive metadata. The new model only has four because Intellectual Entities are now grouped under Objects. See the Appendix for both data models. An additional change was seen in the way Agents are related to Events, Objects, and Rights.

Intellectual Entities

The PREMIS Editorial Committee (2008, 2015) has defined an Intellectual Entity as a "coherent set of content that is described as a unit" (p.212, p.270) and states that intellectual entities can be made up of other intellectual entities. This can include maps, books, databases, and websites, but can also include things that have other intellectual entities nested within such as a piece of software (2015, p.8). Both Kaplan (2009, p. 9) and Guenther et al. (2016, p.27) point out that Intellectual Entities can be a conceptualized abstraction, and the PREMIS Dictionary (2015) states that "an Intellectual Entity may have one or more digital or non-digital Representations," (p.8).

As mentioned above, a major change was made to this entity with the current release. Initially it was represented separately in the data model, a container class that was made up of smaller objects. With release 3.0,

however, it can now be considered a type of object (Caplan, 2017, p.6). This makes sense given the fact that it can be broken into smaller components, a property it already shared with the Object Entity. Although it is no longer considered a separate entity on the data model (see Appendix A), it is being included here since the PREMIS Dictionary (2015) stated that it can be modeled either inside or outside of PREMIS (p. 9), and many preservation projects are still using 2.0.

Objects

The Objects Entity contains 15 semantic units, but only two of them are required. In PREMIS 3.0 the entities in the Object class belong to one of four categories: Intellectual Entity, Representation, File, and Bitstream. This change allows the four categories to have relationships between them that form a "holistic description of the preservation Object," (Dappert, et al., 2013, p.110). This change also brings the data model in line with RDF, while earlier versions were based on XML. See the Appendix (fig.2) for a diagram of relationships within the Object entity.

Events

The ability to document "proper audit trails" (Wilson, 2010, p.209) is an important aspect of preservation management. The Event Entity allows repository managers to create a ledger of events by documenting key

processes in the lifecycle of a digital object. In addition to three required semantic units, there are four others than can be optionally deployed, including some that allow for free-form notes.

The PREMIS Editorial Committee just added ten new Event terms to its controlled vocabulary in 2017. This recognized the fact that preservation needs have experienced a "shift in the depth of knowledge about these activities" (2017a, para. 6), and an increase in the number of organizations engaging in them. The vocabulary has only been updated twice since 2010. Many of the new terms addressed changes in technology. These included technical metadata such as encryption, forensic feature analysis of bitstreams, and validation. It also included administrative events such as accession, filename changes, and redaction.

Agents

In PREMIS 2.0 only the bare minimum was stored about Agents. There is still only one required semantic unit for agent, although there are far more in the current release. More can be understood by reading what Caplan (2009) wrote that "only the minimum needed for identification," (p.5) is to be used. In the 2017 revision the term "minimum" (Caplan, 2017, p.3) is removed. The increased importance of the Agent entity is seen by the fact that there are now eight semantic units for Agent (2017c, p.9), as opposed to only three in the 2009 edition (p.11). While both versions recognize that

an agent can have multiple roles, PREMIS 2.0 included the role in the Events Entity (Caplan, 2009, p.11), but PREMIS 3.0 references it semantically (PREMIS Editorial Committee, 2017c, p.9). This acknowledges the fact that agents might have rights in addition to roles, and these semantic relationships are possible due to the fact that PREMIS is can now be expressed as RDF (PREMIS Editorial Committee, 2015, p.4).

Rights

This is one of the most interesting features of PREMIS and one that, along with Agents and Events, creates a robust and unique semantic framework for preserving digital items. According to Caplan (2017), a major difference between most preservation strategies and PREMIS is that in PREMIS "the rights entity aggregates information about rights and permissions actions that may be restricted by copyright law to the rights holders" (p.9). Typically most preservation projects just aim to preserve the items and their derivatives in a wholesale fashion (p.9). This compilation of rights and associated permissions allows the repository to "do what it needs to do" (p.10), to fully preserve digital items. This means that descriptive metadata is not used by PREMIS to help indicate what rights might exist. Rather PREMIS declares what types of rights the repository has to exchange and disseminate digital objects.

Interoperability

PREMIS is not meant to be used as a standalone metadata schema. Rather it is intended to work with descriptive metadata schemas, providing a backbone of technical and administrative metadata that is needed for preservation purposes. In particular it seems to work fairly well with Metadata Encoding and Transmission Standard (METS) which can include all of the PREMIS metadata in its administrative metadata section (PREMIS Editorial Committee, 2017a, p.2) PREMIS also recommends packaging PREMIS in METS for creating information packages (p.1). Wilson (2010) cautioned "it is important to understand that PREMIS metadata are not sufficient on their own to ensure that reliable, authentic, and meaningful digital data can be carried over time and space," (p.214). By wrapping PREMIS in another metadata schema, it is possible to create robust digital objects that can be preserved, exchanged, and disseminated.

Conclusion

PREMIS is a well-established metadata schema that continues to improve as it keeps pace with developments in LOD. With its newly increased focus on Agents, Events, and Rights, along with its restructured Object class, PREMIS is capable of meeting most preservation needs. Due to the way it describes Objects, there could be some future applicability to

blockchain technology. It will also likely continue to be used to preserve repository objects, particularly where semantic relationships are needed.

References

Caplan, P. (2009). Understanding PREMIS. Washington, DC: Library of Congress Network Development and MARC Standards Office. Retrieved from https://www.loc.gov/standards/premis/understanding-premis.pdf

Caplan, P. (2017). Understanding PREMIS. Washington, DC: Library of Congress Network Development and MARC Standards Office. Retrieved from https://www.loc.gov/standards/premis/understanding-premis-rev2017.pdf

Revision: PREMIS Editorial Committee (2017).

Coppens, S. et al. (2013). PREMIS OWL: A semantic long-term preservation model. International Journal on Digital Libraries, 15:2-4, 87–101. DOI: https://doi.org/10.1007/s00799-014-0136-9

Guenther et al. (2016) Digital Preservation Metadata for Practitioners: Implementing PREMIS. Springer: Cham, Switzerland.

OCLC & RLG. (2005). Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group. Dublin: OH; Mountain View, CA.

retrieved from https://www.loc.gov/standards/premis/v1/premis-dd_1.0_2005_May.pdf

PREMIS Editorial Committee, Library of Congress. (2008). PREMIS data dictionary for preservation metadata: Version 2.0. Retrieved from http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

PREMIS Editorial Committee, Library of Congress. (2015). PREMIS data dictionary for preservation metadata: Version 3.0. Library of Congress Retrieved from https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

PREMIS Editorial Committee, Library of Congress. (2017a). 2017 Revisions of Events Vocabulary. Retrieved from https://www.loc.gov/standards/premis/events-announcement.html

PREMIS Editorial Committee, Library of Congress. (2017b). Guidelines for Using PREMIS with METS for exchange. Retrieved from https://www.loc.gov/standards/premis/guidelines2017-premismets.pdf

PREMIS Editorial Committee, Library of Congress. (2017c). Guidelines for Using the PREMIS Version 3 OWL Ontology. Retrieved from https://www.loc.gov/standards/premis/ontology/pdf/premis3-owl-guidelines.pdf

PREMIS Editorial Committee, Library of Congress. (2017b). PREMIS 3 OWL Ontology Draft Release. Retrieved from https://www.loc.gov/standards/premis/ontology/ontology3-announcement.html

Library of Congress. (2005). PREMIS Data Dictionary for Preservation Metadata. Retrieved from https://www.loc.gov/standards/premis/v1/index.html

Library of Congress. (2013). PREMIS OWL ontology 2.2 now available. Retrieved from https://www.loc.gov/standards/premis/ontology-announcement.html

Wilson, A. (2010). How much is enough: Metadata for preserving digital data. Journal of Library Metadata 10:2-3, pages 75-78. DOI: https://doi.org/10.1080/19386389.2010.506395

Appendix

Figure 1. Data Model for PREMIS 2.0

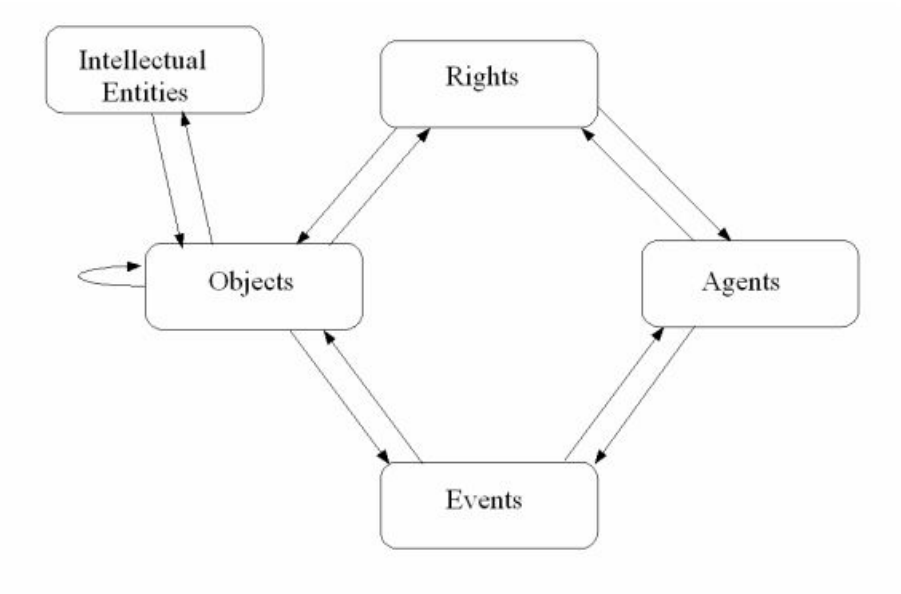
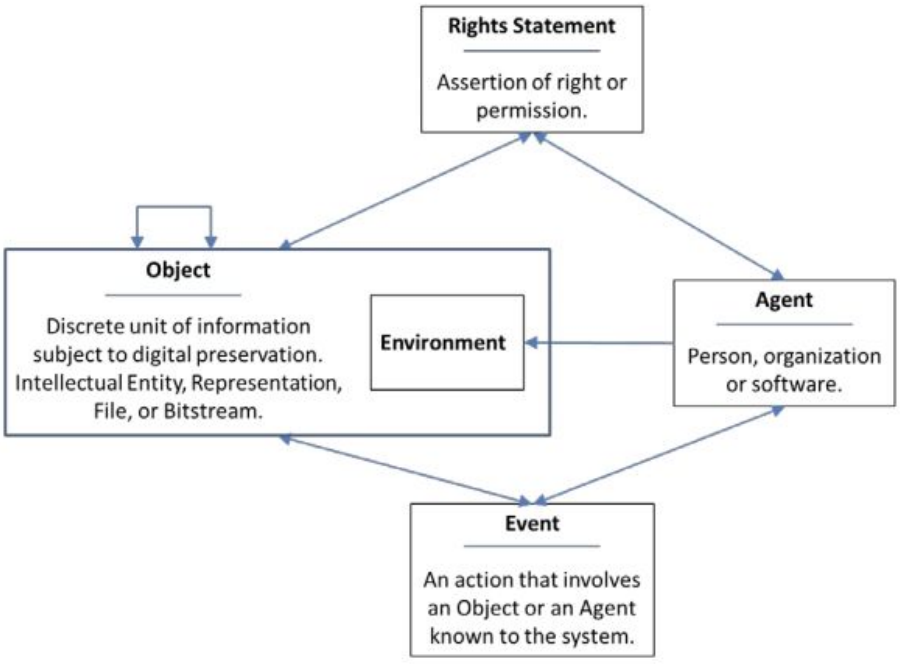


Figure 1A. Data Model for PREMIS 3.0



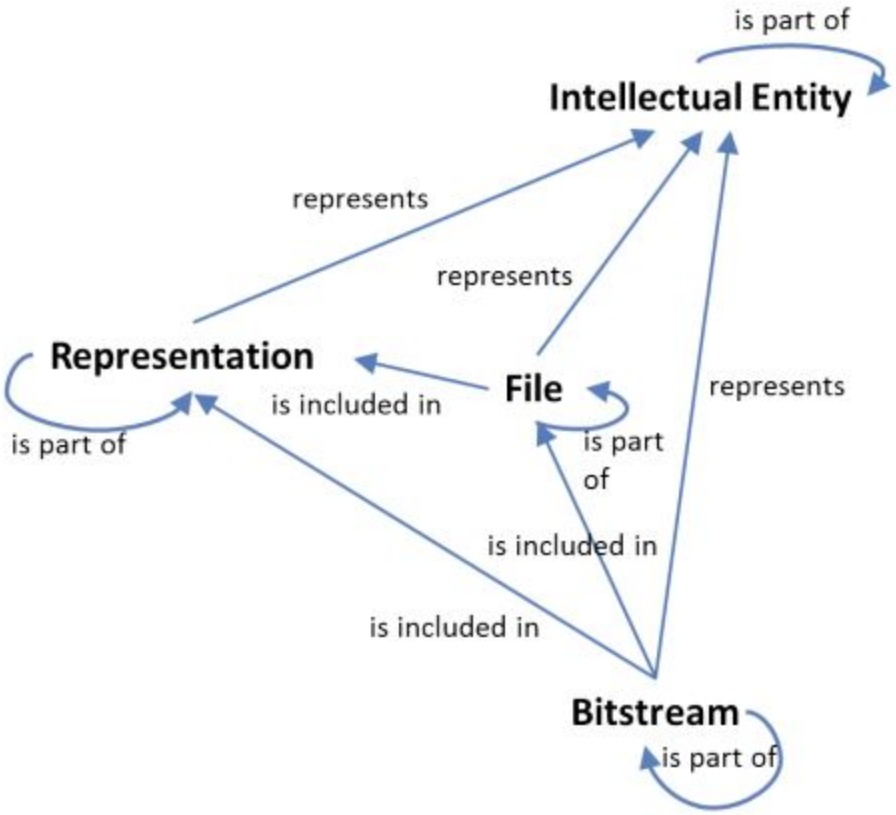


Figure 2: Conceptual view between object categories