PREMIS 3.0 Ontology: Improving Semantic Interoperability of Preservation Metadata

Angela Di Iorio
DIAG - Department of Computer,
Control, and Management
Engineering Antonio Ruberti,
Sapienza University of Rome
Via Ariosto 25 00185, Rome, Italy
angela.diiorio@uniroma1.it

Bertrand Caron
Department of Metadata
Bibliothèque nationale de France
Quai François Mauriac
75706 Paris Cedex 13
bertrand.caron@bnf.fr

ABSTRACT

The PREMIS 3.0 Ontology Working Group is a community interested in using Semantic Web Technology to leverage systems managing the long-term preservation of digital archives.

The version 3 of the PREMIS Data Dictionary has stimulated the community to revise the current PREMIS OWL Ontology. The revision process aims not only to integrate the conceptual model with the changes defined by the new data model of the PREMIS version 3.0, but also to ease the implementation of Semantic Web Technology in the digital preservation community.

Keywords

semantic web technologies; preservation metadata; PREMIS ontology.

1. INTRODUCTION

In this article, the development work for reviewing the PREMIS OWL Ontology [4] is introduced. The PREMIS 3.0 Ontology Working Group is a community interested in using Semantic Web Technology to leverage systems managing the long-term preservation of digital archives.

The current PREMIS OWL is a semantic formalisation of the PREMIS 2.2 Data Dictionary [6] and defines a conceptual model for the metadata that a digital archive needs to know for preserving objects. In June 2015 version 3 of the PREMIS Data Dictionary [7] was released. This in turn has led to a community review of the PREMIS OWL. The review process aims not only to integrate the conceptual model with the changes defined by the data model of the PREMIS version 3.0, but also to ease the implementation of Semantic Web Technology in the digital preservation community.

The PREMIS version 3.0 changed the PREMIS Data Model and refined the description of the digital objects' Environment, a specific type of Intellectual Entity. These changes have implied the revision of the previously published ontology. The revision working group felt that a deeper revision of the existing ontology should be made. Indeed, the previous modelling work had taken as a starting point the PREMIS XML Schema, and automatically transformed it in an OWL file. The obtained ontology was thereby quite close to the Data Dictionary structure and vocabulary, though some simplifications were made to make it more RDF-friendly.

In order to go further in that direction, the PREMIS 3.0 Ontology Working Group decided to look at semantic units of the PREMIS Data Dictionary, not directly as classes and properties, but as description elements of real-world objects. In other words, the dictionary has to be turned into a formalisation of the digital preservation knowledge domain. This perspective implies some significant changes in the ontology. Nevertheless, the revision working group is performing a reconciliation between these necessary changes and the coherence with the PREMIS Data Dictionary.

2. THE EVOLUTION OF THE PREMIS PRESERVATION METADATA

The PREMIS Data Dictionary (PREMIS-DD) is built on the Open Archival Information System (OAIS) reference model (ISO 14721) [2]. The PREMIS-DD defines specifications about which metadata is necessary to preservation practices and provides directions for implementations.

The PREMIS XML schema[1] is usually provided in parallel with the PREMIS-DD for supporting the XML implementation of the preservation metadata management.

The PREMIS Data Model underlying the PREMIS-DD consists of five main information entities [6] deemed important for digital preservation purposes:

1) *Intellectual Entity*, an intellectual unit for the management and the description of the content.

2) *Object*, a discrete unit of information subject to digital preservation. The Object has three subtypes:

    a. *File* is a named and ordered sequence of bytes that is known by an operating system.

    b. *Bitstream* is contiguous or non-contiguous data within a file that has meaningful common properties for preservation purposes.

    c. *Representation* is the set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity.

3) *Event*, an action that has an impact on an Object or an Agent.

4) *Agent*, a person, organization, hardware or software associated with Events in the life of an Object, or with Rights attached to an Object.

5) *Rights*, a description of one or more rights, permissions of an Object or an Agent.

The PREMIS-DD version 3.0 was published in June 2015. The major changes and additions provided by this last version describe dependency relationships between Objects and their Environments: hardware and software needed to use digital objects.

This evolution has required two main repositions:

[1] PREMIS Preservation Metadata XML Schema VERSION 3.0, http://www.loc.gov/standards/premis/premis.xsd

1) the *Intellectual Entity* is defined as a category of Object to enable additional description and linking to related PREMIS entities;
2) the *Environments* (i.e. hardware and software needed to use digital objects) are described as generic Intellectual Entities so that they can be described and preserved reusing the Object entity, as Representation, File or Bitstream.

This change allows for Environment descriptions or even their Representations to be shared. By expanding the scope beyond repository boundaries, the data interoperability among repository systems is improved, because the Environments descriptions is more granular and consistent with their original technological nature.

3. USE CASES AND SCOPE OF THE PREMIS 3.0 ONTOLOGY

The PREMIS 3.0 Ontology Working Group (WG) has initially collected use cases, that would benefit from integrating the use of PREMIS Ontology, in current RDF implementations of digital archives.

The WG has solicited the new version of PREMIS Ontology as a conceptual model for producing RDF datasets expressed in PREMIS 3.0 terms (classes and properties), to be combined with other terms defined by third-party ontologies provided in RDF Schema[2] or OWL [5]. For example, the integration with the Europeana Data Model (EDM)[3], as well as interest in using PREMIS 3.0 Ontology in systems Hydra/Fedora 4 based, by integrating it in the Portland Common Data Model (PCDM)[4], has been discussed by the group and has been considered a feasible test bed for releasing the new ontology.

The general assumption of the WG was that the objective of adopting as much as possible an approach oriented toward the interoperability with other well established ontologies would generally contribute to increase the interoperability of digital archives aiming to use the PREMIS 3.0 Ontology.

Other ontologies that will be considered by the WG for the integration are: PROV-O [12] for provenance information, FOAF[5] for human agents, DOAP[6] for software agents, Dublin Core for descriptive metadata, and OAI-ORE for structural relationships.

Over and above this specific goal (aiming to improve the metadata interoperability of digital repositories) a specific need for improving the interoperability of the management of preservation metadata has also arisen from the WG.

Current practices in searching resources with specific characteristics, usually rely on domain knowledge, expertise, and professional networks of categories, involved in the digital preservation. The cross-repository search can also be complicated by the interoperability problems due to different underlying data models of digital repositories. The need for developing a model connecting different RDF datasets, related to the preservation metadata domain, has led the WG to revise and integrate the current PREMIS OWL ontology.

The integration of third-party ontologies will help to overcome these limitations and to engage user communities to deeply use preservation metadata for supporting their research and to help stakeholders in improving the management of preservation metadata.

The scope of the PREMIS 3.0 Ontology is indeed to support the implementation of different Semantic Web Technology through the digital preservation community, and will support these technologies to answer questions that could arise out of the community (users and stakeholders).

The repositories using the PREMIS 3.0 Ontology as conceptual model for RDF datasets, should have the ability of answering questions like: What is the relationship between an Object and another? How many Files is a Representation made of? What is the average number of Files of all Representations? When was a Representation created? How many Representations were ingested after certain date? Which Files are JPEG images? Which Representations contain Files in PDF format?

4. THE PREVIOUS PREMIS ONTOLOGY

Starting from version 2.2, a PREMIS OWL ontology has been made available alongside the PREMIS XML Schema.

The PREMIS OWL ontology is a semantic formalisation of the PREMIS 2.2 data dictionary [6] and defines a conceptual model for the preservation information of a digital archive. The PREMIS OWL ontology has allowed the interested community to express preservation metadata in RDF, by using the conceptual model of the PREMIS-DD, and as such, it can be used to disseminate the preservation information as Linked (Open) Data [1].

The design of the PREMIS OWL [2] has tried to be coherent as much as possible to the PREMIS-DD, aiming at preserving the knowledge model. As such, the structure of the PREMIS-DD semantic units, defined by experts in the domain of the long-term digital preservation, and its translation in the XML schema, have been replicated in the PREMIS OWL.

The PREMIS OWL has addressed the problem of interoperability, deriving from the preservation policies and processes that each digital preservation archive adopts, by using the formalism of the Web Ontology Language (OWL 1) [5] [8]. In addition, 24 preservation vocabularies have been integrated, that are exposed by the Library of Congress Linked Data Service[7] and are provided as SKOS [8][9] preservation vocabularies.

The PREMIS OWL does not replace but rather complements XML in areas where RDF may be better suited, such as querying or publishing preservation metadata, or connecting repository-specific data to externally maintained registries.

At the time of its design, the PREMIS OWL has deviated from the PREMIS 2.1 Data Dictionary trying to reconcile the model differences between the XML schema and the OWL ontology[8] [2].

The principles and design deviations, as well as the OWL implementation choices have been reviewed by the PREMIS 3.0 Ontology Working Group as a starting point for modelling the PREMIS 3.0 Ontology.

[2] RDF Schema 1.1, https://www.w3.org/TR/rdf-schema/

[3] Europeana Data Model Documentation, http://pro.europeana.eu/page/edm-documentation

[4] Hydra and the Portland Common Data Model (PCDM), https://wiki.duraspace.org/pages/viewpage.action?pageId=69011689

[5] Friend of a Friend (FOAF), http://semanticweb.org/wiki/FOAF

[6] Description of a Project (DOAP), https://github.com/edumbill/doap/wiki

[7] Library of Congress LD Service, http://id.loc.gov/vocabulary/preservation.html

[8] Public workspace for PREMIS OWL ontology, http://premisontologypublic.pbworks.com

5. PREMIS 3.0 ONTOLOGY: WORK IN PROGRESS

In order to make the new version of the ontology more compatible with Linked Data Best Practices[9], the WG followed a similar approach to the one adopted for the revision[10] of the Bibframe[11] ontology. The following principles were agreed upon, though on specific points the working group may decide against them. Some of them were already followed in the previous version of the ontology, some others were not and their adoption may bring important changes in the next version.

5.1 Make it Simple

Simplicity is the key to massive adoption; that is why the working group has the objective of making the ontology as simple as possible; but not simpler. Some of the following principles derive from this generic one, which should be kept in mind at any step of the modeling process.

5.2 Use PREMIS-DD as a Knowledge Base

Having a Data model is a real asset when trying to build an ontology: theoretically, it would provide classes and the Data Dictionary properties. In the case of PREMIS, RDF modeling has to consider other concepts which are in the preservation domain (generally existing as semantic containers in the Data Dictionary) but do not appear in the Data Model, e.g., Signature, Outcome, Registry, etc. Thus the ontology cannot be an exact transcription of the PREMIS Data Dictionary in OWL. The WG had to reconcile two opposite directions: sticking to the PREMIS Data Dictionary or introducing conceptual discrepancies with it in order to reflect more faithfully the preservation activities and to respect ontology design principles.

The PREMIS Data Dictionary is built on the principle of technical neutrality. It gives a list of pieces of information to record without any constraint on where and how to record it. According to the PREMIS conformance principles, implementers can store information anywhere, with any structure and any element names, provided that they can establish an exact mapping between their data and PREMIS semantic units. That is why the WG considers scope, concepts, and intent provided by the Data Dictionary, but feels free to differ regarding the names and structure of the ontology.

As said above, the Data Dictionary provides pieces of information, whereas the ontology describes real-world objects and organizes knowledge on these objects. One example is about semantic containers, a mechanism extensively used by PREMIS to group together related pieces of information. Systematically transcribing them into the ontology would create extra levels of indirection and make data processing more difficult. If high-level containers become classes (e.g. the fixity semantic container becomes a `premis:Fixity` class, as the "fixity" is a real-world concept), for semantic containers of lower level (e.g., formatDesignation, which is only used to group the format name and its version). Their existence as classes in the next version of the ontology is still being debated.

5.3 Re-use Pieces of Existing Ontologies

The scope of the ontology – preservation – covers many other domains: technical characteristics of files, software description, cryptographic functions, structural relationships between digital resources, digital signature, etc. Many of these domains have already defined ontologies whose re-use is worth investigating. Re-using existing vocabularies is one of the most important notions of the semantic web and is agreed best practice, as it is saving time not as much for ontology designers but mainly for developers and consumers. Instead of distrusting other ontologies because of their potential evolution, relying on the expertise of their maintainers seems a better option.

This principle is probably the main difference between the new approach and the previous published ontology, in which re-using vocabularies had been avoided to stick to the Data Dictionary semantic units. The following elements are taking into account when examining the relevance of existing vocabularies, which is made case-by-case:

The classes of an ontology should correspond to concepts within that particular knowledge domain – if PREMIS needs elements that are not specific to the preservation domain, it should ideally pick existing elements in another domain model.

In the case of multiple possible existing ontologies, preference should be given to stable, better-known and more frequently used ones.

When considering re-using an external element, its definition must be taken into account, but also its properties, and especially domain and range, as inference processes will deduce the type of the subject and object of a re-used property. Re-use existing ontologies can thus bring more work to the ontologist but it naturally improves interoperability.

5.4 Re-use of LOC-LDS Preservation Vocabularies

Updates to existing preservation vocabularies and integrations of new ones have been performed[12] coherently with the version 3 of the PREMIS-DD and before of the WG ontology revision. Except for the vocabulary related to the Event types which is still under revision gathering the community feedback, 26 vocabularies have been released. For example, an "Environment function type" vocabulary[13] was created to provide URIs and definitions for the most common types of Environments considered by the PREMIS Editorial Committee: hardware peripheral, plugin, chip, operating system, etc.

Some of the preservation vocabularies were included in the previous version of the ontology; for example, Agent roles[14] were declared subproperties of the `premis:hasEventRelatedAgent`. The same solution was foreseen for the new version of the ontology, in order to manage two different update frequencies, as the ontology should be rather stable compared to vocabularies like software types, which are likely to be submitted to frequent changes. Nevertheless, a discrepancy appears between the ontology, whose classes and properties are designating real-world objects, and preservation vocabularies, which are authoritative vocabularies and designate a concept - they are declared as subclasses of `skos:Concept`. Importing preservation vocabularies which are a collection of simple thesauri and use such terms as subclasses of real-world objects, or re-declaring in the ontology classes and properties as

[9] Best Practices for Publishing Linked Data, http://www.w3.org/TR/ld-bp/

[10] Rob Sanderson, Bibframe Analysis, bit.ly/bibframe-analysis

[11] Bibliographic Framework Initiative, https://www.loc.gov/bibframe/

[12] Revised and new preservation vocabularies, http://id.loc.gov/vocabulary/preservation.html

[13] LOC-CDS Environment function type, http://id.loc.gov/vocabulary/preservation/environmentFunctionType

[14] LOC-CDS Agent role in relation to and Event, http://id.loc.gov/vocabulary/preservation/eventRelatedAgentRole

real-world objects designated by these terms, is still a pending question.

5.5 Establish Equivalent Terms

When re-using is not possible, another way of improving vocabularies interoperability is to declare equivalent terms. In the case the direct re-use of an external element is not chosen because of the element being broader or not directly equivalent, linking the PREMIS element to the external one can be done with properties like (from the meaningful to the most lightweight) the OWL equivalentClass or the RDFS subClassOf and seeAlso properties. For example, the PREMIS class for software Environments could be declared a subclass of the DOAP Project class, so that consumers aware of the DOAP ontology can deduce information about PREMIS software Environments.

Using these properties to link PREMIS ontology elements with elements from other existing ontologies was planned in the previous version of the ontology, though it had not been done.

5.6 Use URIs to Identify Things

Identifying a resource on the web is typically done with URIs, as strings do not provide the same assurance about uniqueness. Literals are dead-ends in linked data, as no assertions can be made from them. Consequently, instead of having a list of values to identify the type of any described entity, a best practice is to create URIs for each item inside the list. To achieve this goal, LOC-LDS preservation vocabularies are considered the reference point, because they provide URIs for terms that are commonly needed by implementers and endorsed by the PREMIS Editorial Committee.

The enumeration is not meant to be comprehensive but extensible: if the list is insufficient to some implementers, they can just coin their own URIs, more tailored to their needs, and declare them members of the corresponding list.

5.7 Follow Best Practices Naming

The names of the classes and predicates should follow best practice naming conventions. Element names should be in "CamelCase". Classes should be initial upper case noun phrases (ClassOfThing), predicates should be initial lowercase verb phrases (hasSomeRelationship). Ambiguous names should be avoided: "isPartOf" / "hasPart" is preferable to "part" which does not indicate at first sight which is the part and which is the whole. Final names of the classes and properties to be created in the ontology can be deferred until the end of the process.

This principle has been followed in the previous ontology. Nevertheless, LOC-LDS preservation vocabularies were designed to be used in different technical contexts (XML files, databases, etc.) and thus do not follow this practice (for example, the URI `http://id.loc.gov/vocabulary/preservation/environmentFunctionType/haa`, possibly abbreviated as `envFuncType:haa`, does not satisfy the requirements for the clarity mentioned above).

5.8 Provide Documentation and Guidelines

As the ontology vocabulary can differ on some points with the Data Dictionary semantic units, documenting all ontology elements and providing guidelines for expressing XML structures as RDF triples is absolutely necessary. The maintenance of documentation and guidelines should not be underestimated either.

6. APPROACH AND TOPICS UNDER DISCUSSION

The PREMIS 3.0 Ontology Working Group has selected specific topics on which focusing the revision process and discussing the design of the PREMIS 3.0 Ontology.

In line with the principles adopted, the general approach has been to revise the conceptual connection between the ontology and the concepts expressed by related LOC-LDS controlled vocabularies (see Section 5.4).

Furthermore, some topics have catalysed questions about choices to be made in developing the conceptual model of the Ontology. Below is provided a list of questions arising around topics and that are being discussed by the WG:

Identifiers and URIs: what is the difference between URIs for identifying RDF resources with respect to the Identifiers semantic unit widely used in the PREMIS-DD? And what is the Identifier entity? Do we need an Identifier class given that identifiers in RDF are the URIs that unambiguously designate resources?

Preservation Level: how is preservation level decided? Are there other entities not included in the PREMIS-DD that could help us modelling PreservationLevel, like for example a top-level Preservation Policy class? Are both preservation levels types and preservation level roles subclasses of Policy? Would it be useful to link them to a policy assignment Event to keep track of their change through migrations?

Significant Properties: are significant properties actually globally true features of the object, or are they assigned by different preservation policies? Would it be useful to link them to a policy assignment Event to keep track of their change through migrations?

The values of significant properties appear to be free text. Is this even useful to record, when it is not machine actionable? Could it just be a `premis:note`?

Environment: has the Objects' environment to be re-modelled, based on the changes in the PREMIS-DD version 3.0?

Agent: Is it possible to define the equivalence of the Agent class with the Agent class defined by the PROV-O or FOAF?

7. FUTURE DEVELOPMENTS

The answers to the listed questions and the choices made for the design of the ontology will lead to the release of the PREMIS 3.0 Ontology. The publication of the new version of the ontology will take into account the provision of proper documentation, following the principles established by the WG.

In addition, the engagement of a wider community, by providing different serialization formats for allowing a wider re-use in the semantic web community will be also considered: the OWL 2 [11] RDF/XML serialization will be released for being used by conforming OWL 2 tools[15]. Additional formats, more readable by the implementers like the Turtle[16] or OWL 2 Functional syntax[17] will be also provided.

8. ACKNOWLEDGMENTS

We would like to thank all members of the PREMIS 3.0 Ontology WG that participated to the group activities and contributed at different level to this article. We thank Rebecca Guenther, Robert Sanderson, Esmé Cowles, Charles Blair, Karen

[15] OWL 2 serialization technical requirements, https://www.w3.org/TR/owl2-overview#Syntaxes

[16] RDF 1.1 Turtle - Terse RDF Triple Language, https://www.w3.org/TR/turtle/

[17] OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition), https://www.w3.org/TR/owl2-syntax/

Estlund, Gloria Gonzalez, Arwen Hutt, Evelyn McLellan, Elizabeth Roke, Ayla Stein, Peter McKinney, Ben Fino-Radin.

REFERENCES

[1] T. Berners-Lee. 2006. Linked data-design issues.

[2] Consultative Committee for Space Data. 2012. Reference Model for an Open Archival Information System (OAIS), Recommended Practice CCSDS 650.0-M-2 Magenta Book.

[3] S. Coppens, S. Peyrard, R. Guenther, K. Ford, T. Creighton. 2011. PREMIS OWL: Introduction, Implementation Guidelines & Best Practices.

[4] S. Coppens, R. Verborgh, S. Peyrard, K. Ford, T. Creighton, R. Guenther, E. Mannens, R. Walle. 2015. Premis owl. Int. J. Digit. Libr., 15(2-4):87-101.

[5] D. L. McGuinness, F. Van Harmelen, et al. Owl web ontology language overview. 2004. W3C recommendation, 10(10):2004.

[6] PREMIS Editorial Committee. 2012. PREMIS Data Dictionary for Preservation Metadata version 2.2.

[7] PREMIS Editorial Committee. 2015. PREMIS Data Dictionary for Preservation Metadata version 3.0.

[8] W3C. 2004. OWL Web Ontology Language Overview.

[9] W3C. 2009. SKOS Simple Knowledge Organization System Primer.

[10] W3C. 2009. SKOS Simple Knowledge Organization System Reference.

[11] W3C. 2012. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition).

[12] W3C. 2013. PROV-O: The PROV Ontology.