

# Digital Long-Term Preservation using a Layered Semantic Metadata Schema of PREMIS 2.0

Sam Coppens, Erik Mannens, Tom Evens, Laurence Hauttekeete & Rik Van de Walle  
Ghent University (Multimedia Lab / MICT) – IBBT  
sam.coppens@ugent.be, erik.mannens@ugent.be, tom.evens@ugent.be,  
laurence.hauttekeete@ugent.be & rik.vandewalle@ugent.be

## Abstract

In Belgium, many institutions have a lot of information stored on analogue carriers. This information is likely to get lost if no digitized copy of the information is stored for the long term. Long-term preservation is subjected to many risks. Overcoming those risks starts with describing the data thoroughly. The metadata needed for long-term preservation are descriptive metadata to search and manage the whole archive, binary metadata to describe the bitstreams, technical metadata describing the files, structural metadata for the representation information, preservation metadata for keeping track of the provenance of the data, and rights metadata. Therefore, we developed a layered semantic metadata schema. The top layer holds the descriptive metadata, the bottom layer holds all the information necessary for long-term preservation. The top layer consist of an OWL representation of Dublin Core, while for the bottom layer we developed an OWL representation of the preservation standard PREMIS 2.0, extended with a vocabulary defining the legal roles of a person, organization, or software. This way, our model offers all the necessary metadata for long-term preservation.

**Keywords:** digital preservation, PREMIS 2.0, ontology, semantic web

## 1 Introduction

In Belgium, the broadcasters, cultural organizations, private persons, and government institutions possess thousands of hours of speech and image material which is stored on analogue carriers. This material belongs to the most important cultural heritage in Flanders. At this moment, the analogue carriers are degrading and are continuously losing quality, making the data inaccessible. Disseminating and storing the content digitally overcomes this problem only temporarily. Furthermore, this digital content has to remain intact and accessible over time, e.g., 20, 50 years or longer. Digital long-term preservation forms the solution for this issue. The project BOM-VI (Preservation and Disclosure of Multimedia Data in Flanders, [1]) initiates the digital long-term preservation of the cultural heritage in Flanders and researches the problems encountered with digital long-term preservation.

In this paper, we present our layered semantic metadata model. First, in chapter two, we introduce the different kinds of metadata that are needed to overcome all the risks involved in long-term preservation and show how our proposed, layered, semantic metadata model relates to those risks. The semantic model consists of two layers: the top layer delivers the descriptive metadata, and the bottom layer is responsible for the binary metadata, the technical metadata, the structural metadata, the preservation metadata (provenance metadata, fixity metadata, and context metadata), and the rights metadata. This way, all the metadata for describing the content for the long-term, are covered by the layered semantic metadata model. For the top layer, we use a Web Ontology Language (OWL, [2]) representation of

Dublin Core [3], which is described in chapter three. For the bottom layer, depicted in chapter four, we developed an OWL representation of the preservation standard Preservation Metadata, Implementation Strategies 2.0 (PREMIS 2.0, [4]). This PREMIS OWL schema (PREMIS OWL, [5]) not only covers the necessary metadata described in chapter two, but also stores the semantics of the metadata for the long term. This can be very important due to, e.g., terminology changes. This schema is accompanied by a vocabulary describing the legal roles that a person, organization, or software application can have.

## 2 Metadata levels for long-term preservation

When preserving digital multimedia data for the long term, the digital archive demands some specific requirements. On the one hand the software and hardware of the digital archive have to guarantee access to the information during a long time. On the other hand human input is necessary in the form of archive descriptions, work processes, and the use of standards to keep the information accessible and interpretable as long as possible to the user community. Based on the Open Archival Information System (OAIS, [6]) reference model, the data has to be described on three levels to guarantee long-term preservation. On every level, there are possible risks for loss of data, which can be minimized by describing the data thoroughly.

On the lowest level, a digital file consists of bits and bytes which can change by external influences, like corruption of carriers, migrations, etc. On this lowest level, **binary metadata** and **fixity metadata** are needed to correct these errors and to guarantee authenticity of the data.

On a higher level, file formats and compression formats like AVI, MP3, and JPEG describe the way the bits can be transformed to an interpretable multimedia representation. When a file format becomes obsolete, the archive has two solutions to preserve the stored data: migration or emulation. Metadata is needed to support these actions. At this level, it is also very important to preserve the look and feel of the objects. when migrating file formats. Thus, a rich description of the look and feel is also necessary. For this level we need **technical metadata**, for describing the files, **structural metadata**, for describing sets of files and their relations, e.g., a book which is represented as a set of scanned TIFF images, and **provenance metadata**, for describing the history of the content information: the original owners of the data, the processes that determined the current form of the data, and the available versions.

On the highest level, the information should remain interpretable. Institution structures, terminologies, the designated community, and the rights of an object or institution can change over time. To keep the information interpretable, enough information should be included in the archived package. At this level, the archive needs **descriptive metadata**, for a general description of the object, e.g., MARC, **rights metadata**, for describing copyright statements, licenses, and possible grants that are given, and **context metadata**, for describing the relations of the content information to information which is not packed in the information package. Examples of context metadata are related datasets, references to documents in the original environment at the moment of publication, helper files, and the language.

When developing a metadata schema for the long-term preservation of digital multimedia, metadata descriptions on all levels have to be taken into account, going from bit level descriptions to descriptions of the intellectual content. To realize this, we developed a layered semantic metadata schema. The top layer offers the descriptive metadata. The bottom layer takes care of the preservation metadata, rights metadata, binary metadata, technical metadata,

and structural metadata necessary for deep archiving. For the top layer, an OWL representation of Dublin Core is developed. For the bottom layer, an OWL representation of the preservation standard PREMIS 2.0 is developed. This standard is based on the OAIS reference model. This schema describes the data on all necessary levels.

### **3 Top layer: Dublin Core**

Descriptive metadata describes the content of the data: subject, author, date of creation, file format, etc. This metadata makes it possible to manage and search the complete digital archive. When archiving data coming from different sectors like the broadcast sector, the libraries, the cultural sector, and the archival sector, a problem arises concerning descriptive metadata. Many of the institutions already have descriptive metadata. Are these descriptive metadata stored as metadata or as data? Both strategies have their advantages and disadvantages. When archiving these descriptions as metadata, the archive has to provide a metadata schema. The choice of this schema is a non-trivial task. The metadata schemes used for the descriptions are very domain-specific. To store the descriptive metadata lossless the descriptive metadata schema should be some kind of smallest common multiple of all the descriptive metadata schemes offered by the institutions. This would be a huge metadata schema, impossible to maintain. That is why the descriptive metadata is archived along with the data in their original metadata format, e.g., MARC , so there is no information loss. On top of this metadata, the archive offers a broadly accepted descriptive metadata schema. This gives the archive the necessary tools to search the whole archive. When finding the data of interest, the original metadata that is stored as data can still be presented to the users.

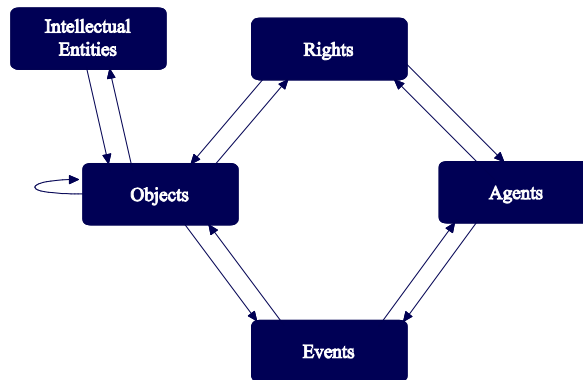
Dublin Core was chosen to describe this top layer of descriptive metadata. Dublin Core is a broadly accepted descriptive schema. The power of this schema is its simplicity and generality. It consists of fifteen fields among which creator, subject, coverage, description, date. It can answer to the basic questions: Who, What, Where, and When. All the fields in Dublin Core are optional and repeatable. This makes it possible to map relatively easily almost all the descriptive metadata schemes to Dublin Core whereas many institutions already support Dublin Core.

### **4 Bottom Layer: PREMIS OWL**

For this layer, we developed an OWL schema of the preservation standard PREMIS 2.0. PREMIS is a preservation standard based on the OAIS reference model. The preservation standard is described by a data model. The data model of PREMIS consists of five semantic units or classes important for digital preservation purposes:

- *Intellectual Entities*: a part of the content that can be considered as an intellectual unit for the management and the description of the content. This can be for example a book, a photo, or a database.
- *Object*: a discrete unit of information in digital form.
- *Event*: An action that has an impact on an object or agent.
- *Agent*: a person, institution, or software application that is related to an event of an object or is associated to the rights of an object.
- *Rights*: description of one or more rights, permissions of an object or agent.

Intellectual entities, events, and rights are directly related to an object. An agent can only be related to an object through an event or through rights. This way, not only the changes to an object are stored, the event involved in this change is also described. These relationships offer the necessary tools to store the provenance of an object properly. Fig. 1 clarifies the data model of PREMIS.



*Fig. 1. Data model of PREMIS*

#### 4.1 Object

The *Object* class describes a unit of information in digital form. It is related to the *intellectual entity* class. This intellectual entity is described by descriptive metadata. This descriptive metadata are very domain-specific. For this, there exist already a lot of descriptive metadata models. Therefore, the description of the intellectual entity is out of scope for PREMIS. In our implementation, the top layer describes the intellectual entity.

An *Object* class knows three subclasses:

- *File*: a file is an ordered sequence of bytes that is known by the system.
- *Bitstream*: a bitstream is the actual data inside a file.
- *Representation*: a representation is a set of files with structural metadata needed for a complete description of an intellectual entity.

The *Object* class possesses all the necessary features to describe the object on the different levels. The minimum information for describing an object (*File*, *Bitstream*, or *Representation*) are *objectIdentifier*, which gives the identifier of the object, *objectCharacteristics*, needed for the *Bitstream* subclass and the *File* subclass, which gives the necessary technical and binary metadata, and *storage*, necessary for describing a *File* or *Bitstream*, which indicates either the location the data is stored, either the medium the data is stored on. An object can be described further into detail using *preservationLevel*, because some repositories offer the opportunity to define a preservation level for an object, *significantProperties*, defining some significant properties of the object, which need to be preserved when, e.g., migrating the data, *originalName*, for indicating the original names of the packages delivered to the repository, *environment*, which describes the environment the user needs to render the content and interact with the content, *signatureInformation*, for storing digital signatures generated during ingest into the repository, and finally, *relationship*, which relates to structural metadata to assemble complex objects.

For linking object information to events, intellectual entities, or rights statements, the object class offers three properties, i.e., *linkingEvent*, *linkingIntellectualEntity*, and *linkingRightsStatement*.

## 4.2 Event

An event aggregates all the information about an action that involves one or more objects. This metadata is stored separately from the object metadata. Actions that modify objects should always be recorded as events.

The *Event* class is described at least by an *eventIdentifier*, *eventType*, e.g. capture, creation, and an *eventDateTime*. This information can be extended using the *eventDetail* property, which gives a more detailed description of the event, and the *eventOutcomeInformation*, which describes the outcome of the event, in terms of success, failure, or partial success. These properties are able to describe any event altering an object. The *Event* class can be related to an *Agent* class or *Object* class via the resp. properties *linkingAgent* and *linkingObject*.

## 4.3 Agent

This class aggregates information about attributes or characteristics of agents. Agents can be persons, organizations or software. This class provides the necessary tools to identify unambiguously an agent. The minimum properties needed to describe the *Agent* class are *agentIdentifier* and *agentType*. Optionally, an agent can also be described using the *agentName*. This is just enough to identify the agent.

An agent can hold or grant one or more rights. It may carry out, authorize, or compel one or more events. An agent can only create or alter an object through an event or with respect to a rights statement. The relationships between an agent and an object through an event or rights entity make it possible to describe the whole provenance of an object.

## 4.4 Rights

The minimum core rights information that a preservation repository must know, is what rights or permissions a repository has to carry out related to objects within the repository. These may be granted by copyright law, by statute, or by a license agreement with the rights holder. Rights entities can be related to one or more objects and one or more agents.

Every *Rights* class can be related to different *RightsStatements*. A *RightsStatement* knows three subclasses: the *Copyright* subclass, the *License* subclass, and the *Statute* subclass. These three subclasses offer the necessary metadata for describing, rights information, i.e., copyrights, licenses, and statutes. Every *RightsStatement* is described at least by a *rightsStatementIdentifier*, and has also the optional property *rightsGranted*, which describes the actions the granting agency has allowed the repository. The *RightsStatement* class can be related to an *Object* class or *Agent* class via the optional, repeatable object properties: *linkingObject* and *linkingAgent*.

This part of the PREMIS OWL schema is extended with a vocabulary that describes the roles agents can have concerning a rights statement. This vocabulary is based on the results of research performed within the project BOM-VI. To fully describe the rights of an object, all the persons, involved in the production of the described object, should be taken into account which is for many organizations impossible. Therefore, a checklist was made with the most important rights and rights holders that should be described. Based on this checklist a vocabulary was made to describe these important legal roles of an agent, e.g., author, composer, conductor.

## 5 Conclusion

When preserving digital information for the long term, different metadata are important. Descriptive metadata are needed to describe the intellectual entities, binary metadata, technical metadata, and structural metadata are essential for the description of the data on all levels (bitstream, file, representation). Preservation metadata is necessary to describe the provenance of the data, to guarantee the authenticity of the digital data, and to provide a context. At last, rights metadata also needs to be stored.

The two-layered, semantic metadata schema described in this paper offers all these metadata. The top layer takes care of the descriptive metadata. An OWL representation of DC was chosen for this layer. The bottom layer carries the binary metadata, technical metadata, structural metadata, preservation metadata, and the rights metadata. For this layer an OWL representation of PREMIS 2.0 was developed. To describe the rights in a more detailed manner, the PREMIS OWL schema was extended with a vocabulary defining the different legal roles of persons, organizations and software. By describing the data with this layered metadata schema, all the risks that come with long-term preservation are minimized. By splitting up the semantic schema in two layers, the top layer with the descriptive metadata can be made public and weaved into the web of data, if the rights permit it. The bottom layer remains closed for the public and is responsible for the long-term preservation of the data.

## 6 References

1. Preservation and Disclosure of Multimedia Data in Flanders, <https://projects.ibbt.be/bom-vl/>
2. Dean, M., Connolly, D., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., Stein, L. A.: OWL web ontology language reference, W3C Working Draft, <http://www.w3.org/TR/2003/WD-owl-ref-20030331> (2003)
3. The Dublin Core Metadata Initiative, DCMI, <http://dublincore.org/> (2009)
4. Higgins, S.: PREMIS Data Dictionary, Digital Curation Centre (DCC), <http://www.loc.gov/standards/premis/v2/premis-dd-2-0.pdf>, Glasgow (2007)
5. Coppens S., Mannens E., Van de Walle R.: PREMIS OWL, Semantic Model of PREMIS 2.0, <http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl>
6. Consultative Committee for Space Data Systems (CCSDS): Reference Model for an Open Archival Information System. Blue book. Issue 1, 148 p., CCSDS, Washington (2002)