briefing paper

digital preservation europe

A data model for preservation metadata

In June 2003 OCLC and RLG established an international PREMIS (Preservation Metadata: Implementation Strategies) working group (WG) which would run for 2 years. The PREMIS WG was composed of more than 30 experts from 5 countries, representing libraries, museums, archives, government agencies, and the private sector. The goal was to define an implementable "set of 'core' preservation metadata elements" for the digital preservation community. In May 2005 a final report, including a data model for preservation metadata and a data dictionary version 1.0, was published. At present the implementation activity is supported by PREMIS maintenance group that is responsible for the schema and data dictionary maintenance and revisions.

Requirements

PREMIS WG has identified five major areas relevant to preservation metadata:

• Provenance: it should record information bearing on the custodial history of the digital object, from the time of the object's creation, and moving forward through successive changes in physical custody and/or ownership.

• Authenticity: it should include information sufficient to validate that the archived digital object is in fact what it purports to be, and that it has not been altered, either intentionally or unintentionally, in an undocumented way.

• Preservation activity: it should document the actions taken over time to preserve the digital object, and record any consequences of these actions that impact the look, feel, or functionality of the object.

• Technical environment: it should describe hardware, operating system, and software applications, needed to render and use the digital object in the state in which it is currently stored in the repository.

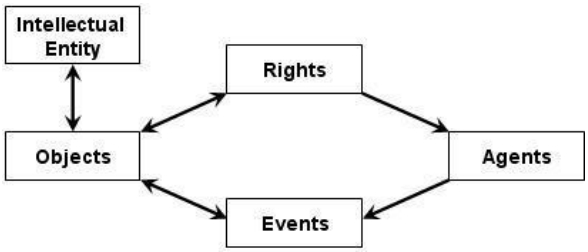
• Rights management: it should record any binding intellectual property rights that limit the repository's powers to take action to preserve the digital object, and to disseminate the object to current and future users.

Data Model summary

The PREMIS data model consists of entities, relationships and properties, that are called semantic units.

Entities

• Intellectual Entity - a coherent set of content that is reasonably described as a unit, for example, a particular book, map, photograph, or database. Because this entity is well described by the descriptive metadata, it is considered out of scope by data dictionary.



• Object or Digital Object - a discrete unit of information in digital form.

• Event - an action that involves at least one object or agent known to the preservation repository.

• Agent - a person, organization, or software program associated with preservation events in the life of an object.

• Rights - assertions of one or more rights or permissions pertaining to an object and/or agent.

Relationships

The relationships are statements of association between instances of entities. "Relationship" can be interpreted broadly or narrowly, and any relationship fact can be expressed in many different ways. The Relationships among Objects appear to be variants of three basic types:

• Structural relationships show relationships between parts of objects. The structural relationships between the files that constitute a representation of an Intellectual Entity are clearly essential preservation metadata. If a preservation repository can not put the pieces of a digital object back together it hasn't preserved the object.

• Derivation relationships result from the replication or transformation of an Object. The intellectual content of the resulting Object is the same, but the Object's instantiation, and possibly its format, are different. Many digital objects are complex, and both structural and derivation information can change over time as a result of preservation activities.
• Dependency relationship exists when one object requires another to support its function, delivery, or coherence of content. The supporting object could not be formally part of the object itself but is necessary to render it.

Properties

The Semantic units are the properties of an entity. In some cases a semantic unit can be a container that groups a set of related semantic units and the grouped subunits are called semantic components of the semantic unit.

The 1:1 principle

The 1:1 principle in metadata asserts that each description describes one and only one resource. As applied to PREMIS metadata, every Object held within the preservation repository (file, bitstream, representation) is described as a static set of bits. It is not possible to modify this set but only to create a new one that is related to the source Object with a derivative relationship. Indeed the Data Dictionary has a semantic unit only for the modification date of an Object because an Object, by definition, cannot be modified.

From the model to dictionary

The data model has an associated data dictionary which includes all the relevant semantic units describing the four entities covered by the dictionary (Objects, Agents, Events, Rights).
In the data model the Relationships among entities of different types are showed by arrows. The Data Dictionary expresses them as linking information, including in the information for entity A a pointer to the related entity B. Every entity in the data model has a unique identifier for use as a pointer. For example, the Object entity has arrows pointing to Intellectual Entities and Events. These are implemented in the Data Dictionary by the semantic units linkingIntellectualEntityIdentifier and linkingEventIdentifier.

Pros and Cons

The PREMIS data dictionary is the result of an international, cross-domain, consensus building process, that can enhance the chance to be widely applicable across all sorts of institutions, digital preservation contexts, and system implementations. As a preservation metadata schema, it attempts to provide comprehensive cover for current needs, directed at practical implementation and interoperability to facilitate objects' transactions, serving its most important function: to document digital objects over time the making them accessible for the long-term. The downside to the PREMIS data dictionary is that nobody can say for certain if its effectiveness will be lasting. The impact of future developments in objects' uses are difficult to foresee.

Conclusions

The Data Dictionary supplies a critical piece of the digital preservation infrastructure, and is a building block with which effective, sustainable digital preservation strategies can be implemented.

Further information and resources

PREMIS (PREservation Metadata: Implementation Strategies) Resources.
http://www.oclc.org/research/projects/pmwg/resources.htm

Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group.
http://www.oclc.org/research/projects/pmwg/premis-final.pdf

Brian Lavoie, Richard Gartner, Preservation Metadata. DPC Technology Watch Report No. 05-01: September 2005
http://www.dpconline.org/docs/reports/dpctw05-01.pdf

PREMIS Working Group.
http://www.oclc.org/research/projects/pmwg/

PREMIS Maintenance Activity.
http://www.loc.gov/standards/premis/

Angela Di Iorio *Fondazione Rinascimento Digitale*
angeladiiorio@gmail.com