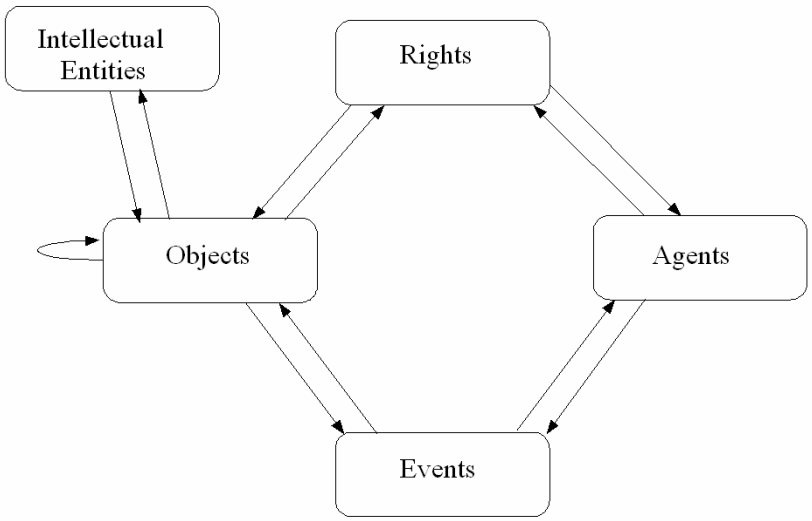
PREMIS [Preservation Metadata Implementation Strategies] Data Dictionary for Preservation Metadata *http://www.loc.gov/standards/premis/*

The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems. The PREMIS Editorial Committee coordinates revisions and implementation of the standard, which consists of the Data Dictionary, an XML schema, and supporting documentation.

PREMIS builds on the OAIS reference model and defines *implementable, core preservation metadata*. PREMIS is implementation independent – that is, it does not concern itself with how metadata is stored, but only with what a repository should know – and therefore defines "semantic units" (metadata that a repository should know/have) rather than "metadata elements" (specific ways of encoding metadata).

Preservation Metadata: "The information a repository uses to support the digital preservation process."

Core Preservation Metadata: Metadata that most repositories will be likely to need to know in order to perform digital preservation. "Core" is not necessarily the same as "mandatory."



In the PREMIS Data Model, boxes represent entities and arrows represent relationships between entities, with the direction of the arrow indicating the direction of the relationship as recorded in the metadata.

Structural Conventions

Semantic Unit: A property of an entity – not to be confused with a metadata element, which is how the information embodied in a semantic unit is actually recorded. Semantic units have *values*. If a semantic unit can apply equally to different entities, it is associated with only one type of entity, and links between the entities are used to clarify the relationship between the semantic unit and other entities.

Mandatory Semantic Unit: Metadata that the repository must know, even if it does not explicitly record the information. "When exchanging PREMIS-conformant metadata with another repository, values for mandatory semantic units must always be provided." "Mandatory" semantic units are actually "mandatory if applicable"; for instance, if the repository does not manage data at the bitstream level, it does not need to use "mandatory" semantic units pertaining to bitstreams.

Container: A semantic unit that is "used to group other related Semantic Units." The grouped semantic units are referred to as **semantic components** of the container. A container does not have a value of its own. Semantic components may themselves be containers.

Extension Container: A container that is defined so as "to allow the use of metadata encoded according to an external schema."

Relationship: "A statement of associations between instances of entities."

Structural Relationship: A relationship between parts of objects.

Derivation Relationship: A relationship between objects where one object is the result of replicating or transforming another. "The intellectual content of the resulting Object is the same, but the Object's instantiation, and possibly its format, are different."

Dependency Relationship: A relationship where "one object requires another to support its function, delivery, or coherence of content." In the Data Dictionary, dependency relationships are part of environment information rather than relationship information.

Identifier: A container whose semantic components uniquely identify an instance of an Object, Event, Agent, or Rights entity. "Identifiers are used as references to establish relationships between entities in the PREMIS data model."

Entities: Sets of "things" that are or can be "described by the same properties".

Intellectual Entity: "A set of content that is considered a single intellectual unit... for example, a particular book, map, photograph, or database. An Intellectual Entity can include other Intellectual Entities" and "may have one or more digital representations." The Data Dictionary does not include semantic units for this Entity "because it is well served by descriptive metadata."

[Digital] Object: "A discrete unit of information in digital form." (This definition is different from the definition used by the digital library community, in which a digital object is "a combination of identifier, metadata, and data.") Object entities have three subtypes:

File: "A named and ordered sequence of bytes that is known by an operating system." Files have file formats, access permissions, and properties like size and last modification date. An "Object consisting of a single File" is a **simple object**.

Bitstream: A set of contiguous or non-contiguous bits in a file that have "meaningful common properties for preservation purposes. A bitstream cannot be transformed into a standalone file without the addition of file structure (headers, etc.) and/or reformatting the bitstream to comply with some particular file format." Bitstreams that can be made into standalone files without adding file structure data or being reformatted are **filestreams**, or "true files

embedded within larger files." In the Data Dictionary, filestreams are considered as more like files than like bitstreams.

Representation: A Digital Object that embodies or instantiates an Intellectual Entity; the group of files, including structural metadata, that make up a version of an intellectual entity so that it can be rendered. "Not all preservation repositories will be concerned with representations", and if a repository does not handle representations, it does not need to record metadata associated with them.

Event: An action that affects or involves at least one Object or Agent. The repository decides which actions it wants to record as Events.

Agent: A "person, organization, or software program/system" that is associated with Events or Rights associated with an Object. The Data Dictionary only defines "a means to identify the agent and a classification of agent type (person, organization, or software)", leaving the definition of other relevant metadata for other initiatives.

Rights: Assertions regarding rights or permissions relating to an Object and/or Agent. This entity is meant "to allow a preservation repository to determine whether it has the right to perform a certain [preservation] action in an automated fashion, with some documentation of the basis for the assertion" and is not generally concerned with rights pertaining to access or distribution.

Other Relevant Definitions and Concepts

Format: "A specific, pre-established structure for the organization of a digital file or bitstream."

Environment: "Digital materials are distinctly different from analog materials because a complex technical environment is interposed between user and content. ... Separating digital content from its environmental context can make the content unusable." The contents of the environment container are the semantic units that the PREMIS working group feels are most essential for a repository to know about an archived object's environment.

The "Onion Model": This model provides a way of thinking about the format of compressed or encrypted objects. It is important to be able to accurately describe layers of encodings and encryptions on an archived object in order to be able to correctly extract the original object. The PREMIS working group "arrived at the metaphor of an onion: a digital object can be wrapped in layers of encodings that need to be 'peeled off' in a particular sequence. The onion model is implemented by treating each layer as a 'composition level,' and organizing metadata into sets of values pertaining to each layer."

Fixity: The property that an object has when it remains unchanged over time.

Integrity: Ensured by identifying and validating the format of a file.

Authenticity: "The quality of [an object] being what it purports to be." If an object is authentic, "the integrity of both [its] source and [its] content ... can be verified."

Digital Provenance: "Documentation of processes in a Digital Object's life cycle." Digital Provenance usually describes key Agents responsible for an Object, important Events occurring during the Object's life cycle, and any other information related to the Object's creation, management, and preservation.