

An Introduction to PREMIS

Jenn Riley

Metadata Librarian

IU Digital Library Program



Outline

- Background and context
- PREMIS data model
- PREMIS data dictionary
- Implementing PREMIS
- Adoption and ongoing developments

What is preservation metadata?

- Information that “supports and documents the digital preservation process”
- The problem is, we don’t really know what that metadata looks like
- Well, we know *something*, but not nearly enough
- The community is thinking hard about the issues, but we still have very little real-world data

Some related work

- National Library of Australia Preservation Metadata for Digital Collections (October 1999)
- Reference Model for an Open Archival Information System (OAIS) (June 2001)
- RLG/OCLC report Trusted Digital Repositories: Attributes and Responsibilities (May 2002)
- OCLC/RLG Metadata Framework to Support the Preservation of Digital Objects (June 2002)
- National Library of New Zealand Metadata Standards Framework –Preservation Metadata (November 2002)
- RLG/NARA Audit Checklist for the Certification of Trusted Digital Repositories (August 2005)

What is PREMIS?

- PREservation Metadata Implementation Strategies
- A working group of over 30 members sponsored by OCLC and RLG
- A data dictionary for preservation metadata included in the May 2005 final report of the working group

Charge to the PREMIS working group

- define an implementable set of “core” preservation metadata elements, with broad applicability within the digital preservation community;
- draft a Data Dictionary to support the core preservation metadata element set;
- examine and evaluate alternative strategies for the encoding, storage, and management of preservation metadata within a digital preservation system, as well as for the exchange of preservation metadata among systems;
- conduct pilot programs for testing the group’s recommendations and best practices in a variety of systems settings; and
- explore opportunities for the cooperative creation and sharing of preservation metadata.

Working group structure

- Implementation Strategies Subgroup
 - “examined various strategies for encoding, storing, and managing preservation metadata within digital preservation systems”
 - performed survey of existing and planned digital preservation systems
- Core Elements Subgroup
 - defined core elements
 - drafted data dictionary

How PREMIS defines preservation metadata

- “The information a repository uses to support the digital preservation **process**”
- Metadata that supports
 - viability
 - renderability
 - understandability
 - authenticity
 - identity
- Mandatory elements represent “the minimum amount for [a] second repository to accept custody of [a] digital object and assume responsibility for its long-term preservation”

PREMIS goals

- Build on the OAIS reference model
- Be implementation independent
- “Provide a starting point for improvements and enhancements based on community experience and feedback”

Development strategies

- Paid particular attention to documenting
 - digital provenance
 - relationships
- “Whenever possible the group defined elements that do not require human intervention to supply or analyze,” but did not limit to these
- Defined “semantic units” rather than “metadata elements”

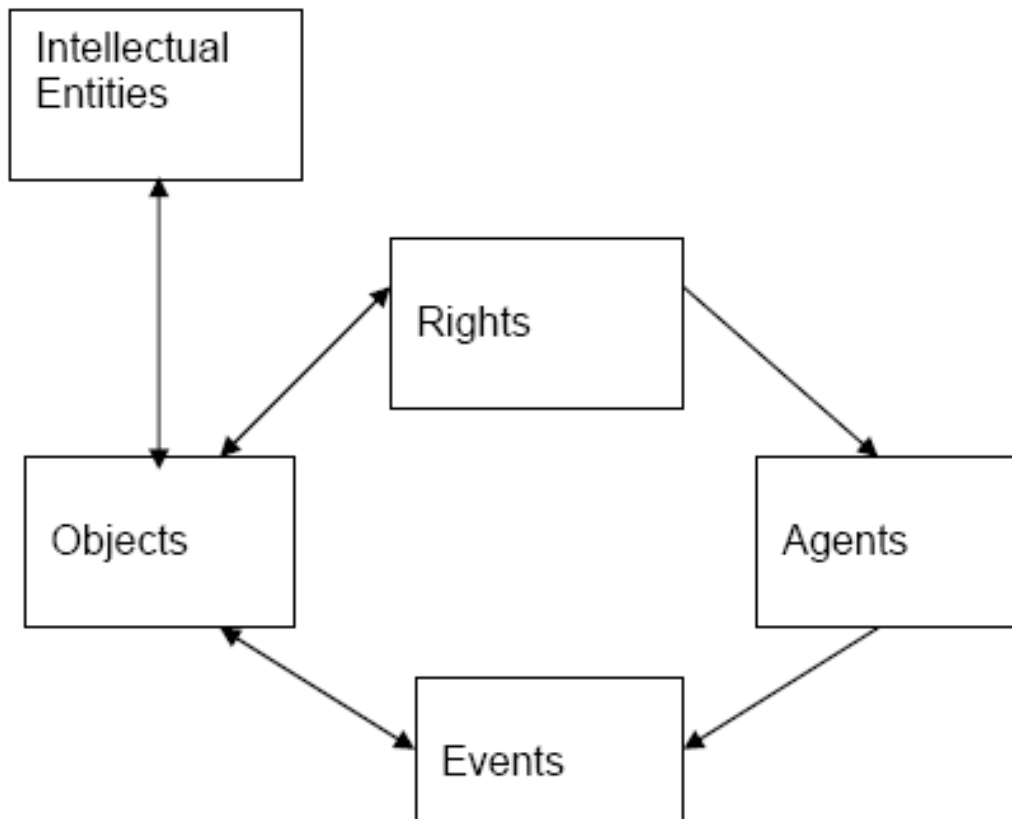
Defining “core”

- “Things that most working preservation repositories are likely to need to know in order to support digital preservation”
- “Core does not necessarily mean mandatory”
- “Core elements define information that a repository needs to know, regardless of how, or even whether, that information is stored”
- Core elements support checking of:
 - Fixity – object is unchanged since some previous time
 - Integrity – compliant with relevant specifications
 - Authenticity – object is what it purports to be

Outline

- Background and context
- PREMIS data model
- PREMIS data dictionary
- Implementing PREMIS
- Adoption and ongoing developments

The PREMIS data model



Intellectual entities

- “A coherent set of content that is reasonably described as a unit”
- Can include other Intellectual Entities
- May have one or more digital representations
- May not be managed by all repositories

Objects

- “A discrete unit of information in digital form”
- Is a static set of bits that cannot be modified
- Three subtypes
 - File
 - Bitstream
 - Representation

File object

- “A named and ordered sequence of bytes that is known by an operating system”
- Defined like “file” in common usage
- No restriction on format, etc.

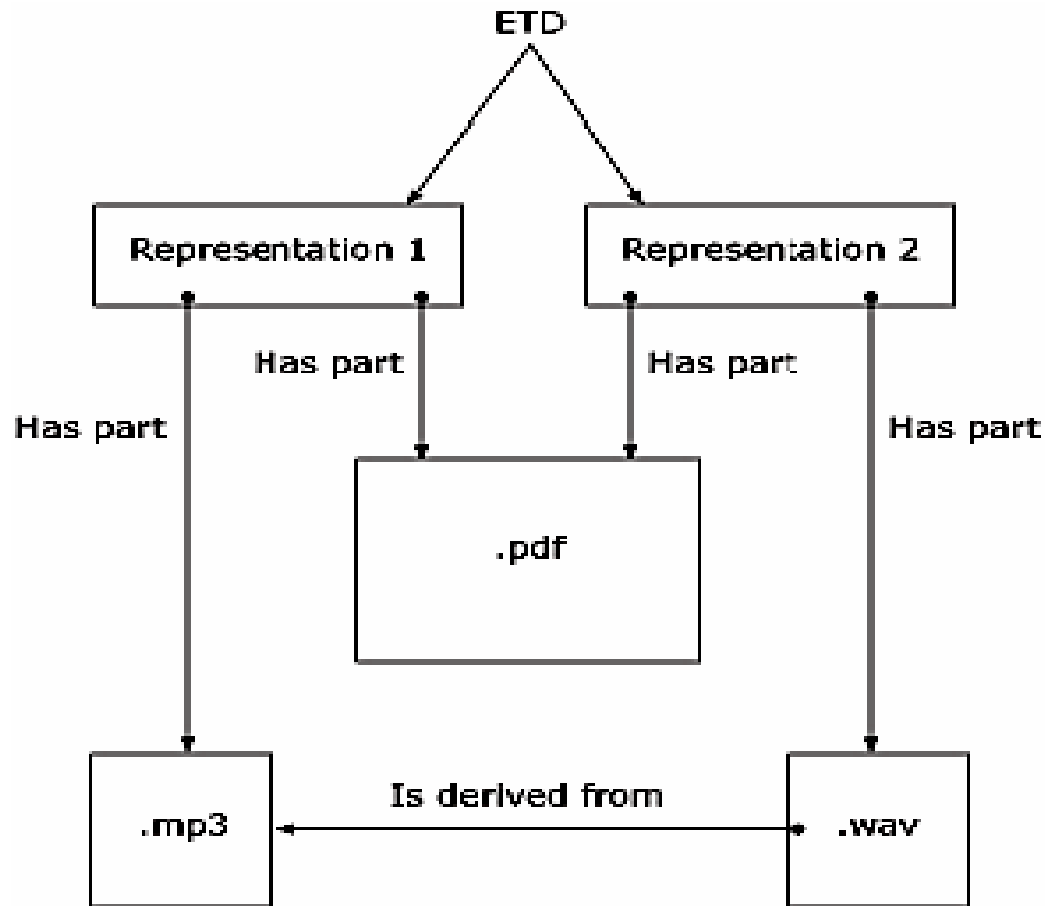
Bitstream object

- “Contiguous or non-contiguous data within a file that has meaningful common properties for preservation purposes”
- Defined differently than common usage
 - Can’t span files
 - Must have some sort of reformatting to be made into a file
- *Weird exception: filestreams*
 - Bitstreams that don’t need additional information to be transformed into a file
 - Follow all the file rules in the data dictionary, not the bitstream rules

Representation object

- “The set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity”
- More than one Representation may exist for each Intellectual Entity
- Repository doesn’t necessarily have to track representations

Objects example: ETD



Events

- The Event entity aggregates metadata about actions that involve at least one object or agent known to the preservation repository
- Many types of Events might be of interest to a preservation repository
 - Creation of a new version of an object
 - Create/alter relationships
 - Validity/integrity checking
 - etc.
- All Events have outcomes
- Some Events have outputs

Agents

- “A person, organization, or software program associated with preservation events in the life of an object”
- Agents represented minimally in PREMIS
 - Means of identification
 - Classification as person, organization, or software
- Assumes other initiatives will more fully define Agents
- Agents influence Objects only indirectly through Events

Rights

- Rights Statements “are assertions of one or more rights or permissions pertaining to an Object and/or Agent”
- Semantic units related to rights restricted to those concerned with preservation activities
- All expressible as “Agent A grants this *permission* for Object B.”
- 3 semantic units
 - allowed act
 - expiration date of the permission
 - all other terms, conditions, restrictions and/or limitations
- Acknowledges much more work needs to be done

Relationships between entities

- Between objects
 - Structural relationships
 - Derivation relationships
 - Dependency relationships
- Others defined by data model indicated in data dictionary by linking attributes

Outline

- Background and context
- PREMIS data model
- PREMIS data dictionary
- Implementing PREMIS
- Adoption and ongoing developments

The PREMIS data dictionary

- Defines semantic units for:
 - Objects
 - Events
 - Agents
 - Rights
- Intellectual Entity is out of scope because it is “well served by descriptive metadata”

Entries include information on:

- Name
- Semantic components
- Definition
- Rationale
- Data constraint
- Object category
- Applicability
- Examples
- Repeatability
- Obligation
- Creation/Maintenance notes
- Usage notes

Sample data dictionary entry

Semantic unit	format		
Semantic components	formatDesignation, formatRegistry		
Definition	Identification of the format of a file or bitstream where format is the organization of digital information according to preset specifications.		
Rationale	Many preservation activities depend on detailed knowledge about the format of the digital object. An accurate identification of format is essential. The identification provided, whether by name or pointer into a format registry, should be sufficient to associate the object with more detailed format information.		
Data constraint	Container		
Object category	Representation	File	Bitstream
Applicability	Not applicable	Applicable	Applicable
Repeatability		Not repeatable	Not repeatable
Obligation		Mandatory	Mandatory
Creation/ Maintenance notes	The format of a file or bitstream should be ascertained by the repository on ingest. Even if this information is provided by the submitter, directly in metadata or indirectly via the file name extension, recommended practice is to independently identify the format by parsing the file when possible. If the format can not be identified at the time of ingest, it is valid to record that the format is unknown, but the repository should subsequently make an effort to identify the format, even if manual intervention is required.		
Usage notes	<p>A bitstream embedded within a file may have different characteristics than the larger file. For example, a bitstream in LaTeX format could be embedded within an SGML file, or multiple images using different colorspaces could be embedded within a TIFF file. Format must be recorded for every file. When the bitstream format can be recognized by the repository and the repository might want to treat the bitstream differently from the embedding file for preservation purposes, format can be recorded for embedded bitstreams.</p> <p>Either formatDesignation or formatRegistry should be recorded. Both are optional, but since format (the container) is mandatory, one of these must be used.</p> <p>See "Format information," page 4-1.</p>		

Outline

- Background and context
- PREMIS data model
- PREMIS data dictionary
- Implementing PREMIS
- Adoption and ongoing developments

Role of a preservation policy

- PREMIS helps a repository to implement a preservation policy; it doesn't *set* that policy
- Policy can be complicated
 - Is descriptive metadata part of an Intellectual Entity?
 - If so, should we treat it as a file?
 - Is PREMIS data itself a file (or a bitstream) that is managed by the repository?
 - etc., ad infinitum...
- The data dictionary is only a starting point, does not include all information needed to preserve an Object

Relationship to technical metadata

- PREMIS semantic units restricted to:
 - intellectual characteristics
 - characteristics common to all formats
- Some overlap between PREMIS semantic units and elements defined by various technical metadata standards

Other types of metadata

- Structural and rights metadata fall at least partly within the scope of PREMIS, but perhaps not entirely
- Descriptive metadata is useful for defining Intellectual Entities, which are managed by some repositories

Lack of relevant content standards and controlled vocabularies

- ❑ Only in some cases do definitions of semantic elements provide guidance on how to structure the data recorded
- ❑ PREMIS semantic units largely outside the scope of most existing content standards
- ❑ “PREMIS assumes that repositories will adopt or define controlled vocabularies useful to them”
- ❑ Perhaps a common content standard isn't needed, but the lack of one does mean more decisions have to be made when implementing a repository

PREMIS conformance

- “Local metadata can be used to extend but not modify the PREMIS semantic units”
- “The mandatory semantic units of the Data Dictionary represent the information that a preservation repository must be able to associate with any archived digital object in its possession”
- Don’t have to *store* PREMIS information, just have to *know* it
- Currently no formal means of stating or testing “conformance”

XML Schemas

- Literal representations of the semantic units and attributes of the PREMIS data dictionary
- Of use for exchange of preservation objects
- Likely of less use for a repository's internal representation
- 5 separate schemas
 - PREMIS container
 - Object entity
 - Event entity
 - Agent entity
 - Rights entity



Outline

- Background and context
- PREMIS data model
- PREMIS data dictionary
- Implementing PREMIS
- Adoption and ongoing developments

Adoption

- Hard to tell, since preservation repositories operate behind the scenes
- PREMIS Implementation Registry currently has 8 diverse entries
- Active implementer's discussion list
- Working Group won the 2005 Digital Preservation Coalition Digital Preservation Award
- Several PREMIS workshops scheduled

Current PREMIS activity

- PREMIS Maintenance Activity hosted at the Library of Congress
- Editorial Committee named
- Commissioned report on Rights in the PREMIS Data Model
- Proposals for revisions of two semantic units in public comment period

So where are we?

- The data dictionary looks to be having a big impact
- Discussion of preservation metadata is increasing
- It seems the PREMIS goal of a “starting point” has been well fulfilled
- PREMIS looks promising as a source of ideas for the IU DLP preservation repository

For more information

- PREMIS Working Group site:
<<http://www.oclc.org/research/projects/pmwg/default.htm>>
- PREMIS Maintenance Activity site:
<<http://www.loc.gov/standards/premis/>>
- PREMIS Implementors Listserv:
<<http://listserv.loc.gov/listarch/pig.html>>
- These presentation slides:
<<http://www.dlib.indiana.edu/~jenlrile/presentations/bbspr07/premis/premis.ppt>>
- jenlrile@indiana.edu