The Promise of PREMIS:
background, scope and purpose of the Data Dictionary for Preservation Metadata

Rebecca Guenther, Library of Congress

Long-term Repositories: taking the shock out of the future
Aug. 31-Sept. 1, 2006
Sponsored by APSR
OUTLINE

- Background
  - What is preservation metadata
  - Early work in preservation metadata
- PREMIS charge and scope
- PREMIS data model
- The PREMIS data dictionary
- Implementation issues
- PREMIS maintenance activity
Preservation metadata includes:

- **Provenance:**
  - *Who has had custody/ownership of the digital object?*

- **Authenticity:**
  - *Is the digital object what it purports to be?*

- **Preservation Activity:**
  - *What has been done to preserve the digital object?*

- **Technical Environment:**
  - *What is needed to render and use the digital object?*

- **Rights Management:**
  - *What IPR must be observed?*

➢ **Makes digital objects self-documenting across time**
Early work in preservation metadata

- Open Archival Information System (OAIS)
  - defined a basic abstract information model
- NLA, CEDARS and NEDLIB
  - developed preservation metadata schemes for their projects
  - unified earlier work within the OAIS framework
- National Library of New Zealand, 2002
  - organized metadata elements around a data model
- Preservation Metadata: Implementation Strategies (PREMIS)
  - focused on practical implementation needs
From theory to practice ...

Preservation Metadata Requirements

OAIS → Framework → PREMIS Data Dictionary → Digital Archiving Systems
PREMIS Working Group

Objective: Define implementable, core preservation metadata, with recommendations for management and use

Membership:
- > 30 experts from 5 countries, libraries, museums, archives, government agencies, private sector
- Co-Chairs: Priscilla Caplan (FCLA), Rebecca Guenther (LC)

Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group:
- PREMIS Data Dictionary 1.0
- Accompanying report (scope, context, data model, special topics, glossary, examples)
- XML schemas to support implementation
Some guiding principles and assumptions ...

- “Implementable, core, preservation metadata”:  
  - “Preservation metadata”: maintain viability, renderability, understandability, authenticity, identity in a preservation context  
  - “Core”: What most preservation repositories need to know to preserve digital materials over the long-term  
  - “Implementable”: rigorously defined; supported by usage guidelines/recommendations; emphasis on automated workflows

- Implementation neutral:  
  - No assumptions on specific implementation  
  - Promote flexibility/interoperability  
  - Focus on semantic units: what you need to know (implementation-neutral) vs. metadata elements: how you record it (implementation-specific)  
  - Information that needs to be “recoverable” from the digital archiving system, independent of local implementation
Uses and scope

PREMIS can provide:
- Common data model for organizing/thinking about preservation metadata
- Guidance for local implementations
- Standard for exchanging information packages between repositories

PREMIS is not designed to provide:
- Out-of-the-box solution: need to instantiate as metadata elements in the repository system
- All needed metadata: excludes business rules, format-specific technical metadata, descriptive metadata for access, non-core preservation metadata
- Lifecycle management of objects outside the repository
- Rights management: limited to permissions to perform actions within the repository
Assumes stuff arrives in SIPs and is stored in AIPs, and PREMIS is what the repository needs to know to ingest, store and preserve it for the future.
PREMIS data model
Intellectual Entity

A coherent set of content that is reasonably described as a unit, for example, a particular book, map, photograph, or database.

- May include other Intellectual Entities (e.g. as a website includes a web page).
- May have one or more digital representations.
- Can reference an Object or be referenced by an Object, but is not described in PREMIS.

Examples:
- Rabbit Run by John Updike (a book)
- [Maggie at the beach] (a photograph)
- The Library of Congress Website (a website)
- The Library of Congress: American Memory Home page (a web page)
Object

Examples:
- chapter1.pdf (a pdf file)
- chapter1.pdf + chapter2.pdf + chapter3.pdf (the pdf version of a book in 3 chapters)
- an audio stream in uncompressed pcm (a bitstream within an AVI file)
- a video stream in MJPEG (a bitstream within an AVI file)

- A discrete unit of information in digital form.
- Objects are what the repository preserves.
- **FILE** = a named and ordered sequence of bytes that is known by an operating system.
- **REPRESENTATION** = the set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity.
- **BITSTREAM** = contiguous or non-contiguous data within a file that has meaningful common properties for preservation purposes.
OBJECTS: A book in two versions

IE: Atonement by Ian McEwan

Rep 1: page image version

File 1: page1.tiff
File 2: page2.tiff
File N: pageN.tiff

Rep 2: ebook version

File N+1 METS.xml
File 1: book.lit
Event

- An action that involves at least one object or agent known to the preservation repository.
- Who, what, how, when, and to which object.
- Necessary to document digital provenance. Can track history of object through the events in the object’s life.

Examples:
- A validation event: verifying that chapter1.pdf is a good PDF file
- An ingest event: completing the process of creating an AIP for a SIP
- A migration event: creating a new version of an object in a more contemporary format
Agent

Examples:
- Evan Owens (a person)
- Bank of Scotland (an organization)
- Bank of Scotland, Computer Systems Department (an organization)
- JHOVE version 1.0 (a software program)

- A person, organization, or software program associated with preservation events in the life of an object.
- Not defined in detail in PREMIS; not considered core preservation metadata beyond identification.
Rights

Example:
- The Bank of Scotland gives the repository permission to make an unlimited number of copies of chapter1.pdf under its Agreement with the repository signed December 11, 2006.

- An agreement with a rightsholder that allows a repository to take action(s) related to objects in the repository.
- Not a full rights expression language.
- Assumption the repository is the grantee.
- Basic statement is: Agent A grants Permission P for Object B.
The PREMIS Data Dictionary

Data Dictionary
for Preservation Metadata

Contents:

Acknowledgments
Introduction
The PREMIS Data Model
The PREMIS Data Dictionary version 1.0
Examples
Special Topics
Methodology
Implementation Considerations
Glossary

Final Report of the PREMIS Working Group
May 2005

Preservation Metadata: Implementation Strategies (PREMIS)
A working group jointly sponsored by OCLC and RLG
## Sample Data Dictionary entry

<table>
<thead>
<tr>
<th>Semantic unit</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic components</td>
<td>None</td>
</tr>
<tr>
<td>Definition</td>
<td>The size in bytes of the file or bitstream stored in the repository.</td>
</tr>
<tr>
<td>Rationale</td>
<td>Size is useful for ensuring the correct number of bytes from storage have been retrieved and that an application has enough room to move or process files. It might also be used when billing for storage.</td>
</tr>
<tr>
<td>Data constraint</td>
<td>Integer</td>
</tr>
<tr>
<td>Object category</td>
<td>Representation</td>
</tr>
<tr>
<td>Applicability</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Examples</td>
<td></td>
</tr>
<tr>
<td>Repeatability</td>
<td>Not repeatable</td>
</tr>
<tr>
<td>Obligation</td>
<td>Optional</td>
</tr>
<tr>
<td>Creation/Maintenance notes</td>
<td>Automatically obtained by the repository.</td>
</tr>
<tr>
<td>Usage notes</td>
<td>Defining this semantic unit as size in bytes makes it unnecessary to record a unit of measurement. However, for the purpose of data exchange the unit of measurement should be stated or understood by both partners.</td>
</tr>
</tbody>
</table>
Object entity

- Aggregates characteristics relevant to preservation management that are properties of the object
- Semantic units may not all be applicable to each type of object (representation, file, bitstream)
- Main types of information
  - identifier
  - object characteristics (includes technical properties common to all or most formats)
  - creation information
  - software and hardware environment
  - digital signatures
  - relationships to other objects
  - links to other types of entity
Agents

- The Agent entity aggregates information about agents (persons, organizations, or software) associated with rights management and/or preservation events in the life of an object.
- Intended only to identify the agent unambiguously, and to allow linking from other entity types.
- Repositories encouraged to use any richer scheme that may be appropriate.
- **Semantic units**
  - agentIdentifier (mandatory)
  - agentIdentifierType (mandatory)
  - agentIdentifierValue (mandatory)
  - agentName (optional)
  - agentType (optional)
Events

- The Events entity aggregates information about an action involving one or more Objects
- Recording events can be very important
  - to demonstrate digital provenance
  - to prove that rights have not been violated
  - as an audit trail
  - for problem solving if something goes wrong
  - for billing or reporting
- **Semantic units**
  - eventIdentifier (mandatory)
  - eventType (mandatory)
  - eventDateTime (mandatory)
  - eventDetail (optional)
  - eventOutcomeInformation (optional)
  - linkingAgentIdentifier (optional)
  - linkingObjectIdentifier (optional)
Rights entity

- The Rights entity aggregates information about statements of permissions

- PREMIS addresses only narrow scope: what permissions have been granted to the repository itself to carry out actions related to objects within the repository

- **Semantic units for rights**
  - permissionStatement
    - permissionStatementIdentifier (mandatory)
    - linkingObject (mandatory)
    - grantingAgent (optional)
    - grantingAgreement (optional)
    - permissionGranted (mandatory)
      - act (mandatory)
      - restriction (optional)
      - termOfGrant (mandatory)
      - permissionNote (optional)
Community interest

- As of July 2006:
  - ~25,000 hits on Data Dictionary
  - More than 100 subscribers to the PREMIS Implementers’ Group (PIG) discussion list

- Awarded the U.K. Digital Preservation Award for 2005 and the SAA Preservation Publication Award for 2006

- The PREMIS Data Dictionary is a product of collaboration and consensus
  - Digital preservation is a shared problem which invites shared solutions

- Multiplicity of perspectives on the working group helps promote applicability in many contexts
  - The Data Dictionary should be useful to any institution committed to the long-term preservation of digital materials
Implementation issues

- How PREMIS may be used:
  - For existing repositories (as a checklist for evaluation)
  - For systems in development (as a basis for metadata definition)

- Reconciling data models
  - PREMIS data model is for convenience of aggregation
  - Many arbitrary decisions, e.g. is an anomaly discovered during validation a property of the object or an outcome of the validation event?
  - Other data models equally valid, e.g. NLNZ has Process, Object, File, Metadata
  - However: PREMIS encourages consistent application of preservation metadata across different categories of objects (representation, file, bitstream)

- Implementation in relational databases
  - PREMIS data model is not entity-relationship model
Implementation issues: obtaining values and conformance

- Obtaining values
  - Most can be populated by program but tools would help
    - JHOVE, NLNZ Metadata Extraction Tool
  - Need registries for format and environment information
    - Pronom, GDFR
- What values to use for controlled vocabularies?
  - PREMIS does not have “scheme” element but probably should
- Conformance: defined in PREMIS report
  - local metadata can supplement but not modify PREMIS
  - can define more stringent repeatability and obligation but not more liberal
  - meaning of mandatory:
    - you have to know it, and you have to be able to supply it if exporting for exchange
    - you don’t have to record it in repository
Implementation issues: need for additional metadata

- preservation metadata not considered core
  - core = all objects, all preservation strategies
  - example of non-core = installation requirements

- more detailed information on Rights and Agents

- metadata describing Intellectual Entity

- format-specific technical metadata

- business rules of the repository

- information about the metadata itself (e.g., who obtained or recorded a value, when last changed...
PREMIS XML schemas

- One schema for each PREMIS entity in data model
  - Allows user to choose which parts of PREMIS to use

- PREMIS container schema
  - References schema for each entity type
  - Provides a container if it is desirable to keep some or all PREMIS metadata together
  - If using container requires at least an object which in turn requires objectIdentifier and objectCategory
  - Individual schemas may used alone or with container

- Semantic units in PREMIS schemas
  - XML is faithful to data dictionary
  - Only those units mandatory for all categories of objects are mandatory in object schema
PREMIS in METS: what is METS?

- METS records the (possibly hierarchical) structure of digital objects, the names and locations of the files that comprise those objects, and the associated metadata
- A METS document may be a unit of storage or a transmission format
- METS uses extension “wrappers” or “sockets” where elements from other schemas can be plugged in
- METS uses the XML Schema facility for combining vocabularies from different Namespaces
- The METS Editorial Board has endorsed PREMIS as an extension schema
Main sections of a METS Document

<mets>
  <dmdSec/>
  <amdSec/>
    <techMD/>
    <rightsMD/>
    <sourceMD/>
    <digiProvMD/>
  <fileSec/>
  <structMap/>
</mets>
Issues in using PREMIS with METS

- Which METS sections to use and how many
- Whether to record elements redundantly in PREMIS that are defined explicitly in the METS schema
- How to record elements that are also part of a format specific technical metadata schema (e.g. MIX)
- Recording structural relationships
- How to deal with locally controlled vocabularies
- Whether to use the PREMIS container
- Experimentation will lead to best practices
- An LC example:  http://www.loc.gov/premis/louis.xml
PREMIS Maintenance Activity

Permanent Web presence, hosted by Library of Congress

Centralized destination for information, announcements, and other PREMIS-related resources

Discussion list for PREMIS implementers (PIG list)

Coordinate future revisions of Data Dictionary and XML schema

Editorial committee recently established to guide development and revisions

http://www.loc.gov/standards/premis/
Some implementers ...

- **MathArc (Germany):** A joint project funded by NSF (Cornell) and SUB Göttingen (DFG) to build a distributed archive for mathematical journals distributed between two archives to keep information redundant.

- **DAITTSS (Florida):** a preservation repository for the use of the libraries of the public universities of Florida. Uses a locally-developed software application (DAITSS), which implements most of the PREMIS data elements.

- **Ex Libris (DigiTool):** an enterprise solution for the management of digital assets in libraries and academic environments consisting of a number of modules, each designed to address different needs, functions, and workflows pertaining to the life cycle of a digital object.

- For more information see:  
  - http://www.loc.gov/premis/premis-registry.html
Going forward ...

- Convene new Editorial Committee
- First revision of Data Dictionary and schemas
- Work with other initiatives (e.g., METS, Z39.87) to integrate PREMIS with existing standards, technologies, best practices

Consultancies:
- Rights issues for digital preservation (Karen Coyle)
- PREMIS implementation recommendations (Deborah Woodyard-Robinson)

PREMIS tutorials
- Digital Curation Center PREMIS tutorial (July 17-18 Glasgow)
- DLF tutorial (probably Nov. 2006)
- Other tutorials?