An introduction to PREMIS
Plan

- Background
- Data model and key concepts
- Object
- Event
- Agent
- Rights
- PREMIS evolutions
- Some implementation considerations
Background

- Need for a **common reference** for core preservation metadata:
  - core elements of information
  - guidelines on how they should be recorded

- **2003**: OCLC / RLG **PREMIS working group**
  PREservation metadata:implementation strategies
  Based on the OAIS information model
  Goal: core preservation metadata
    - Data dictionary with implementation guidelines
PREMIS: birth, state-of-the-art and next steps

Before
- May 2005: PREMIS 1.0 Data Dictionary & XML Schema
- March 2008: PREMIS 2.0 Data Dictionary & XML Schema

Now
  This tutorial is based on PREMIS 2.1

What’s next?
- **Oct. 2011**: publication of a draft OWL ontology
  Based on the 2.1 Data Dictionary
- **Coming soon**: PREMIS 3.0 Data Dictionary & XML Schema
What’s in PREMIS?

- "Things" you have to describe
  PREMIS Data model

- What you want to say about these "things"
  PREMIS Data dictionary

- How you want this information to be encoded and implemented
  In XML → PREMIS XML schema
  In RDF → OWL ontology
  Or any other way you like it
The data model: 5 interacting entities

- Intellectual Entity
- Object
- Rights
- Agent
- Event

(identifiers, identifiers, identifiers, identifiers)
From the data model to the data dictionary

- Data model: defines **Entities** and **relationships** between them

- Data Dictionary: for each Entity lists its **semantic units**
  A semantic unit is a property of an entity:
  - Something you *need to know* about an Object, Event, Agent, Right
  - A piece of information most repositories need to know in order to carry out their digital preservation functions

- Two kinds of semantic unit:
  - **Container**: groups together related semantic units
  - **Semantic components**: semantic units grouped under the same container

- Example:
  
  ObjectIdentifier [container]
  
  ObjectIdentifierType [semantic component]
  
  ObjectIdentifierValue [semantic component]
Identifiers in PREMIS

- Identifiers used to
  - **identify** unambiguously an object, agent, event, rights statement...
    - [entity]Identifier
  - and **link** it to another entity
    - linking[entity]Identifier

- All identifiers have
  - An identifierType (category of identifier)
  - An identifierValue (the identifier itself)

- identifierType optimally should contain sufficient information to indicate:
  - How to build the value
  - Who is the naming authority
  - The domain under which the identifier is unique
  - Examples: URL, DOI, ARK, local...

- If all identifiers are local to the repository system, identifierType does not necessarily have to be recorded for each identifier in the system
  - BUT it should be supplied when exchanging data with others
PREMIS identifiers in action

- Intellectual Entity
  - linking IntellectualEntity Identifier
- Object
  - relatedObject Identifier
  - linkingObjectIdentifier
- Rights
  - rightsStatement Identifier
- Event
  - eventIdentifier
- Agent
  - agentIdentifier

XML:
- `<xml />`

Tue, Nov 1st, 2011
Extension containers in PREMIS

- PREMIS is **core preservation metadata**
- PREMIS defines an Extension container to extend PREMIS if you need
  - more granular description
  - specific semantic units (non-core information)
  - out of scope semantic units (not grounded in preservation)
- Extensions are **empty containers**
  - Its semantic components are **whatever you need**
  - One schema per extension; if more schemas are needed, the extension element needs to be repeated
  - Mechanism in PREMIS XML Schema: `<mdSec>` element
- Data in the container may replace, refine or be additional to the appropriate PREMIS semantic unit
3 categories of objects

Objects are what repositories actually preserve

**FILE:** named and ordered sequence of bytes that is known by an operating system

**REPRESENTATION:** set of files that, taken together, constitute a complete rendering of an Intellectual Entity

**BITSTREAM:** data within a file with properties relevant for preservation purposes (but needs additional structure or reformatting to be stand-alone file)

**FILESTREAMS** (files within files) are considered *files* since they can be rendered alone
Intellectual Entities

- Set of content that is considered a single intellectual unit for purposes of management and description (e.g., a book, a photograph, a map, a database)
- Has one or more digital representations
- May include other Intellectual Entities (e.g. a website that includes a web page)
- Not fully described in PREMIS DD, but can be linked to in metadata describing digital representation

Examples:
- “Maggie at the beach” (a photograph)
- The Library of Congress Website (a website)

**THIS WILL CHANGE IN 3.0**
Example: one content, 3 digital representations

Intellectual entity ➔

Representation 2

Files
Part3.pdf Partn.pdf

Representation 1

Files
page1.tif page2.tif page3.tif
page1.xml page2.xml page3.xml

Representation 3

File
Book.epub
Object: high level semantic units

what technical information on it? objectCharacteristics

which object is it? objectIdentifier ark:/12148/btp6k102002g/f1

what kind of object? objectCategory

what is my preservation strategy for this object? preservationLevel

where is it stored? on which media? storage

what software or hardware should be used to handle the object? environment

which of its characteristics do I want to preserve in it? significantProperties

TIFF 10Mb
Object: high level semantic units

objectIdentifier (M,R)
objectCategory (M,NR)
preservationLevel (O,R) [representation, file]
significantProperties (O,R)
objectCharacteristics (M,R) [file, bitstream]
originalName (O,NR)
storage (O,R) [file, bitstream]
environment (O,R)
signatureInformation (O,R) [file, bitstream]
Relationship (O,R)
linkingEventIdentifier (O,R)
linkingIntellectualEntityIdentifier (O,R)
linkingRightsStatementIdentifier (O,R)
Relationships between Objects

- **structural**

  - relationship
    - relationshipType: structural / derivation
    - relationshipSubType: is part of, is source of...
    - relatedObjectIdentification
    - relatedObjectIdentifierType
    - relatedObjectIdentifierValue
    - relatedObjectSequence

- **derivation**

  - is source of
objectCharacteristics [for file or bitstream]

- what checksum?
  - fixity
  - 0a7d048211f3c4dc e3a85c9c89a65651

- what’s its size in bytes?
  - size
  - 15484580

- what format?
  - format

- what application was used to create it?
  - creatingApplication

- do I need to express format specific information?
  - objectCharacteristicsExtension

- access restrictions on this object?
  - inhibitors
    - (password, encryption...)

- is the object directly renderable?
  - compositionLevel
objectCharacteristics [for file or bitstream]

compositionLevel (M, NR)
fixity (O, R)
  messageDigestAlgorithm (M, NR)
  messageDigest (M, NR)
  messageDigestOriginator (O, NR)
size (O, NR)
format (M, R)
creatingApplication (O, R)
  creatingApplicationName (O, NR)
  creatingApplicationVersion (O, NR)
  dateCreatedByApplication (O, NR)
  creatingApplicationExtension (O, R)
inhibitors (O, R)
objectCharacteristicsExtension (O, R)
compositionLevel

sometimes there is more than one layer of characteristics

- compositionLevel = 0
- format = PDF
- size = 500,000 bytes
- messageDigest = [something]

- compositionLevel = 1
- format = gzip
- size = 324,876 bytes
- messageDigest = [something else]
**PREMIS** PREservation Metadata Implementation Strategies

= **different composition Levels**

Number of operations needed to access the primary data object

| composition Level | chapter1.pdf | | chapter1.pdf.gz |
|-------------------|--------------|--------------|
| fixity            | 0            | 1            |
| Message Digest    | SHA-1        | message Digest |
| Algorithm         |              | Algorithm |
| [big string]      |              | [another string] |
| Fixity            | Submitter    | Repository   |
| Message Digest    |              |              |
| Originator        |              |              |
| Size              | 500000       | 324876       |
| format            | format Designation |
| Name              | PDF          |
| format            | format Designation |
| Version           | 1.2          | format Designation |
| version           | 1.2.3        | format Version |
An introduction to PREMIS format

Features:
1. Basic information about the format
2. Link to some more detailed description in a format registry

Semantic units

<table>
<thead>
<tr>
<th>Format</th>
<th>Sample Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>formatDesignation (O,NR)</td>
<td>image/tiff</td>
</tr>
<tr>
<td>formatName (M,NR)</td>
<td>6.0</td>
</tr>
<tr>
<td>formatVersion (O,NR)</td>
<td></td>
</tr>
<tr>
<td>formatRegistry</td>
<td>PRONOM</td>
</tr>
<tr>
<td>formatRegistryName (M,NR)</td>
<td>fmt/353</td>
</tr>
<tr>
<td>formatRegistryKey (M,NR)</td>
<td>format specifications</td>
</tr>
<tr>
<td>formatRegistryRole (O,NR)</td>
<td><a href="http://www.nationalarchives.gov.uk">http://www.nationalarchives.gov.uk</a></td>
</tr>
<tr>
<td>formatNote (O,R)</td>
<td></td>
</tr>
</tbody>
</table>
objectCharacteristicsExtension: an example

<premis:mdSec>
  <premis:mdWrap MDTYPE="TEXTMD" MIME_TYPE="text/xml">
    <premis:xmlData>
      <textmd:textMD xmlns:textmd="info:lc/xmlns/textMD-v3">
        <textmd:character_info>
          <textmd:charset>ISO-8859-1</textmd:charset>
          <textmd:byte_order>little</textmd:byte_order>
          <textmd:byte_size>8</textmd:byte_size>
          <textmd:character_size>1</textmd:character_size>
          <textmd:linebreak>CR/LF</textmd:linebreak>
        </textmd:character_info>
        <textmd:markup_basis version="1.0">XML</textmd:markup_basis>
      </textmd:textMD>
    </premis:xmlData>
  </premis:mdWrap>
</premis:mdSec>
Event examples

- An action that involves or impacts at least one Object or Agent associated with or known by the preservation repository.

- Helps document digital provenance. Can track history of Object through the chain of Events that occur during the Object’s lifecycle.

- Determining which Events are in scope is up to the repository (e.g., Events which occur before ingest, or after de-accession).

- Determining which Events should be recorded, and at what level of granularity is up to the repository.

Examples:
- Validation Event: use JHOVE tool to verify that part1.pdf is a valid PDF file.
- Ingest Event: transform an OAIS SIP into an AIP (one Event or multiple Events?)
Event: high level semantic units

eventIdentifier (M,NR)
eventType (M,NR)
eventDateTime (M,NR)
eventDetail (O,NR)
eventOutcomeInformation (O,R)
linkingAgentIdentifier (O,R)
linkingObjectIdentifier (O,R)
This event has an outcome. it has processed successfully. but how precisely? here is the machine response in plain text. or here is the response in structured fashion.
**eventOutcomeInformation**

```
eventOutcomeInformation
eventOutcome
eventOutcomeDetail
eventOutcomeDetailNote
```

**Sample description**

validation event

```
validation process successful
well-formed and valid
(or)
<Whole XML output of JHOVE>
```
Agent examples

Examples:
- Sébastien Peyrard (a person)
- French national library (an organization)
- JHOVE version 1.5 (a software program)

- Not defined in detail in PREMIS Data Dictionary:
- Not considered core preservation metadata beyond identification
Agent: semantic units

agentIdentifier
agentIdentifierType
agentIdentifierValue
agentName
agentType
agentNote
agentExtension

Sample description

URI
info:bnf/spar/agent/jhove_1_5
JHOVE 1.5
software
Release notes:
http://sourceforge.net/projects/jhove/files/jhove/JHOVE%201.5/RELEASENOTES
Rights statement examples

- An agreement with a rights holder that grants permission for the repository to undertake an action(s) associated with an Object(s) in the repository.

- Not a full rights expression language; focuses on permissions that take the form:
  - Agent X grants Permission Y to the repository in regard to Object Z.

- Basis for rights may be copyright, license or statute.
Rights statement: high level semantic units

rightsStatement
  rightsStatementIdentifier
  rightsBasis
  copyrightInformation
  licenseInformation
  statuteInformation
  rightsGranted
  linkingObjectIdentifier
  linkingAgentIdentifier

rightsExtension

Either rightsStatement or rightsExtension must be present
rightsStatement: 3 possible rights bases

- intellectual property statute
- copyright
- legislation
- statute
- agreement with the rightsholders
- license

What does this mean in the repository?

rightsGranted
rightsBasis → copyright, statute, license

If the basis is copyright, copyrightInformation must be present.
If the basis is license, licenseInformation must be present.
If the basis is statute, then statuteInformation must be present.

rightsStatement
  rightsStatementIdentifier
  rightsBasis © ☢️
  copyrightInformation
  licenseInformation
  statuteInformation
rightsGranted

rightsGranted
act
restriction
termOfGrant
startDate
endDate

what action is allowed?
on which conditions?
from when to when?
rightsGranted

Sample description

rightsholder must be notified

startDate 2010-05-05
endDate 2015-05-04
## 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>2038927</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>
What’s next?

PREMIS OWL ontology
PREMIS 3.0 evolutions
PREMIS OWL ontology in a nutshell

- **Purpose**
  - Providing the community with an RDF serialization of the PREMIS data model and dictionary
  - While remaining as close as possible to the data dictionary’s clearly defined semantics

RDF modelling in 3 words:
- Everything modelled under the form of subject-verb-object
- But what objects? what verbs? what objects?
  - role of vocabularies & ontologies

PREMIS OWL example
PREMIS 3.0: evolution of the data model

Intellectual entities become a category of object
PREMIS 3.0: rights changes (work in progress)

- Ability to declare **other rights bases**, e.g. the policy of a particular institution
  - Addition of an otherRightsInformation semantic element
  - Mechanism: if rightsBasis = other → use otherRightsInformation
- Ability to link to **documentation** supporting some rights statement
- Addition of a termOfRestriction
  - termOfGrant gives the period during which the permissions are granted
  - termOfRestriction gives the time period during which a restriction applies (useful for embargoes)

New in PREMIS 3.0
Implementing PREMIS: toolbox
PREMIS Maintenance Activity

- **Web site:**
  - Permanent Web presence, hosted by Library of Congress
  - Central location for PREMIS-related info, announcements, resources
  - Home of the PREMIS Implementers’ Group (PIG) discussion list

- **PREMIS Editorial Committee:**
  - Set directions/priorities for PREMIS development
  - Considers proposals for changes
  - Coordinates revisions of Data Dictionary and XML schema

http://www.loc.gov/standards/premis/
PREMIS Conformance

- Conformant Implementation of the PREMIS Data Dictionary
- What does "being conformant to PREMIS" mean?
- Conformant at which level?
  - **semantic unit**: conformant implementation of the information defined in a particular semantic unit
  - **data dictionary**: conformant implementation of all semantic units
- Conformant from what perspective?
  - **internal**: conformant implementation at semantic units and data dictionary levels
  - **external** (exchanging PREMIS descriptions):
    - import = the repository can manage PREMIS conformant information
    - export = the repository can provide others with PREMIS conformant information
PREMIS conformance – degrees of freedom

- What am I free to do now?
  - **naming**: using different names from the data dictionary
  - **granularity**:
    - a single metadata element can aggregate semantic units
    - information from a semantic unit can be split in multiple metadata elements
  - **level of detail**: adding more detailed information than the data dictionary
  - **explicit recording of mandatory semantic units**: need not be recorded BUT this information must be recoverable
  - **use of controlled vocabularies**: it is recommended but not mandatory to use controlled vocabularies, defined internally or externally
### Some externally controlled vocabularies

<table>
<thead>
<tr>
<th>Semantic unit</th>
<th>2.2 eventType</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic components</td>
<td>None</td>
</tr>
<tr>
<td>Definition</td>
<td>A categorization of the nature of the event.</td>
</tr>
<tr>
<td>Rationale</td>
<td>Categorizing events will aid the preservation repository in machine processing of event information, particularly in reporting.</td>
</tr>
<tr>
<td>Data constraint</td>
<td>Value should be taken from a controlled vocabulary.</td>
</tr>
<tr>
<td>Examples</td>
<td>E77 [a code used within a repository for a particular event type]</td>
</tr>
<tr>
<td></td>
<td>Ingest</td>
</tr>
<tr>
<td>Repeatability</td>
<td>Not repeatable</td>
</tr>
<tr>
<td>Obligation</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Usage notes</td>
<td>Each repository should define its own controlled vocabulary of eventType values. A suggested starter list for consideration (see also the <a href="#">Glossary</a> for more detailed definitions):</td>
</tr>
</tbody>
</table>
Controlled vocabularies

- Library of Congress is establishing databases with controlled vocabulary values for standards that it maintains
- Controlled lists are represented using SKOS as well as alternative syntaxes
- [http://id.loc.gov](http://id.loc.gov)
- Some lists are relevant for PREMIS:
  - Preservation events
  - Cryptographic hash algorithms
  - Preservation level role
- Will be adding additional PREMIS controlled vocabularies in the near future
Questions?

sebastien.peyrd@bnf.fr