P R E M I S PREservation Metadata Implementation Strategies

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

- Jan. 2011: PREMIS 2.1 Data Dictionary & XML Schema
  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

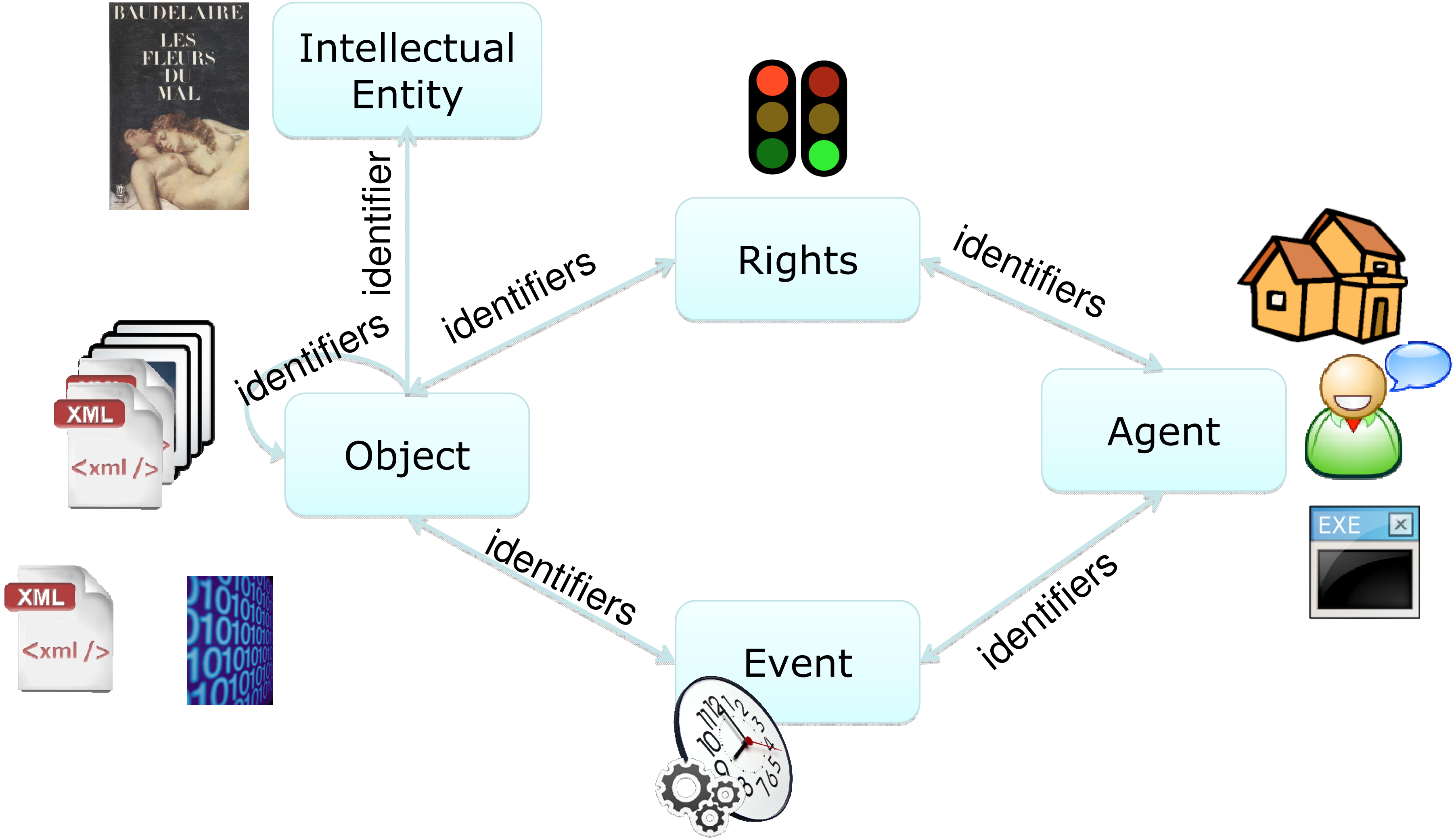
- How you want this information to be encoded and implemented

  In XML → PREMIS XML schema

  In RDF → OWL ontology

  Or any other way you like it

The data model: 5 interacting entities



Intellectual Entity

identifier

Object

identifiers

Rights

identifiers

identifiers

Agent

identifiers

Event

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

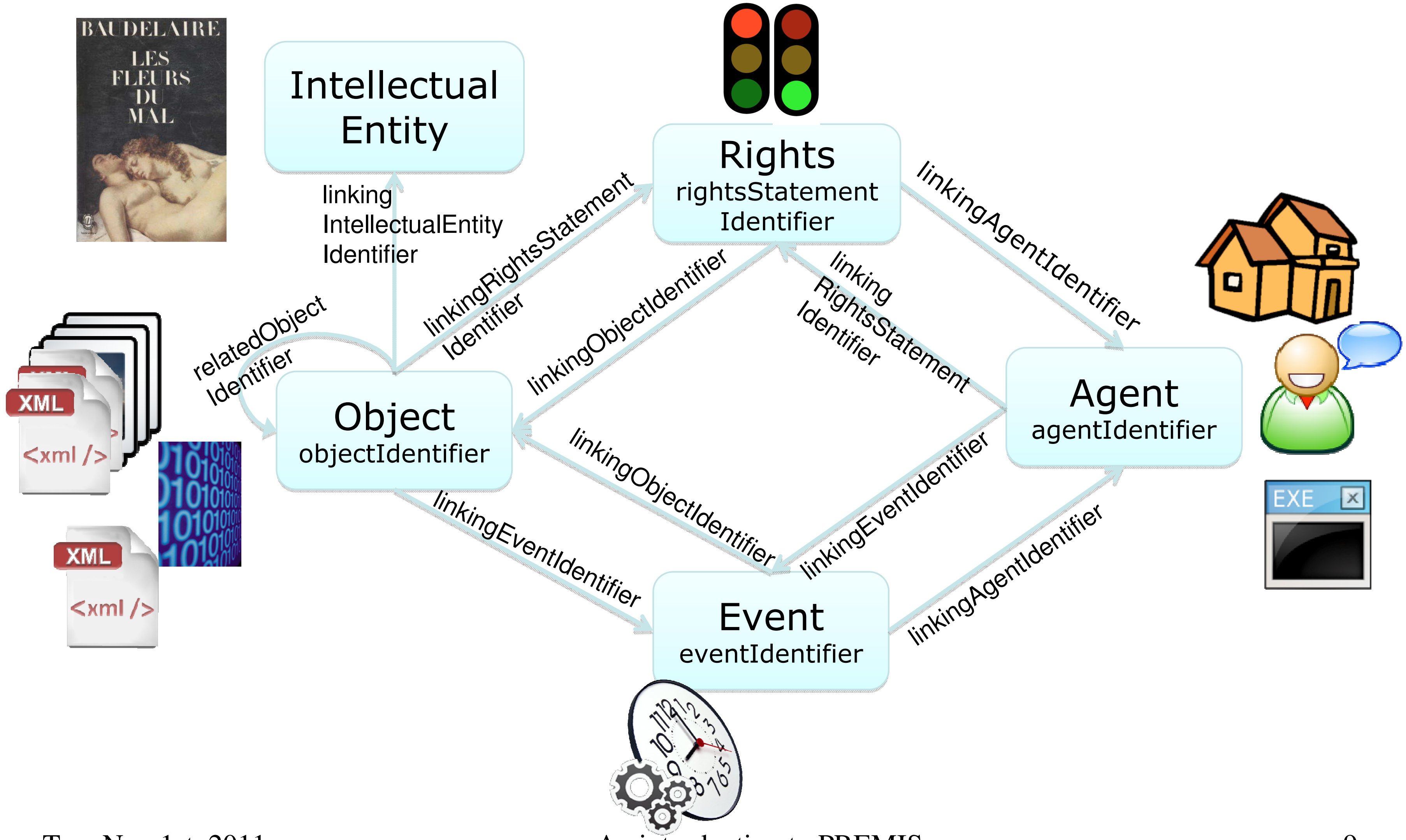
- All identifiers have
  - An identifierType (category of identifier)
  - An identifierValue (the identifier itself)

- 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

Rights rightsStatement Identifier

Object objectIdentifier

Agent agentIdentifier

Event eventIdentifier

relatedObject Identifier

linkingRightsStatement Identifier

linkingObjectIdentifier

linking RightsStatement Identifier

linkingAgentIdentifier

linkingObjectIdentifier

linkingEventIdentifier

linkingEventIdentifier

linkingAgentIdentifier

Extension containers in PREMIS

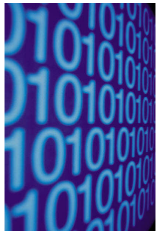
- PREMIS is **core preservation metadata**
- PREMIS defines an Extension container to extend PREMIS if you need
  - more granular description
  - specific semantic units (non-core information)
  - out of scope semantic units (not grounded in preservation)
- Extensions are **empty containers**
  - Its 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

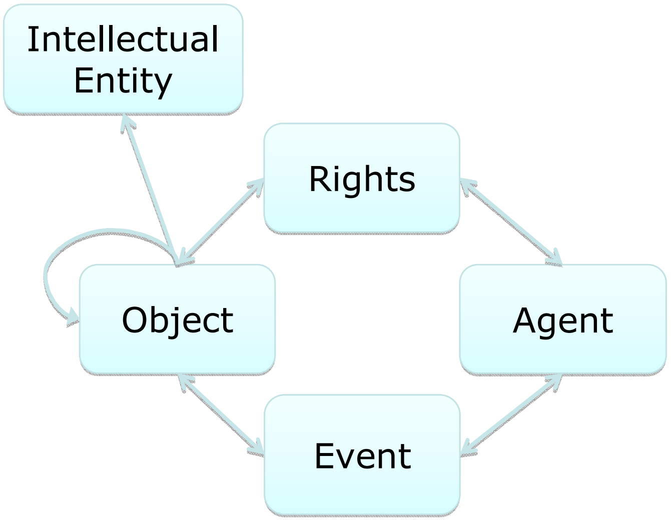
REPRESENTATION: set of files that, taken together, constitute a complete rendering of an Intellectual Entity



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



Intellectual Entity

Rights

Object

Agent

Event

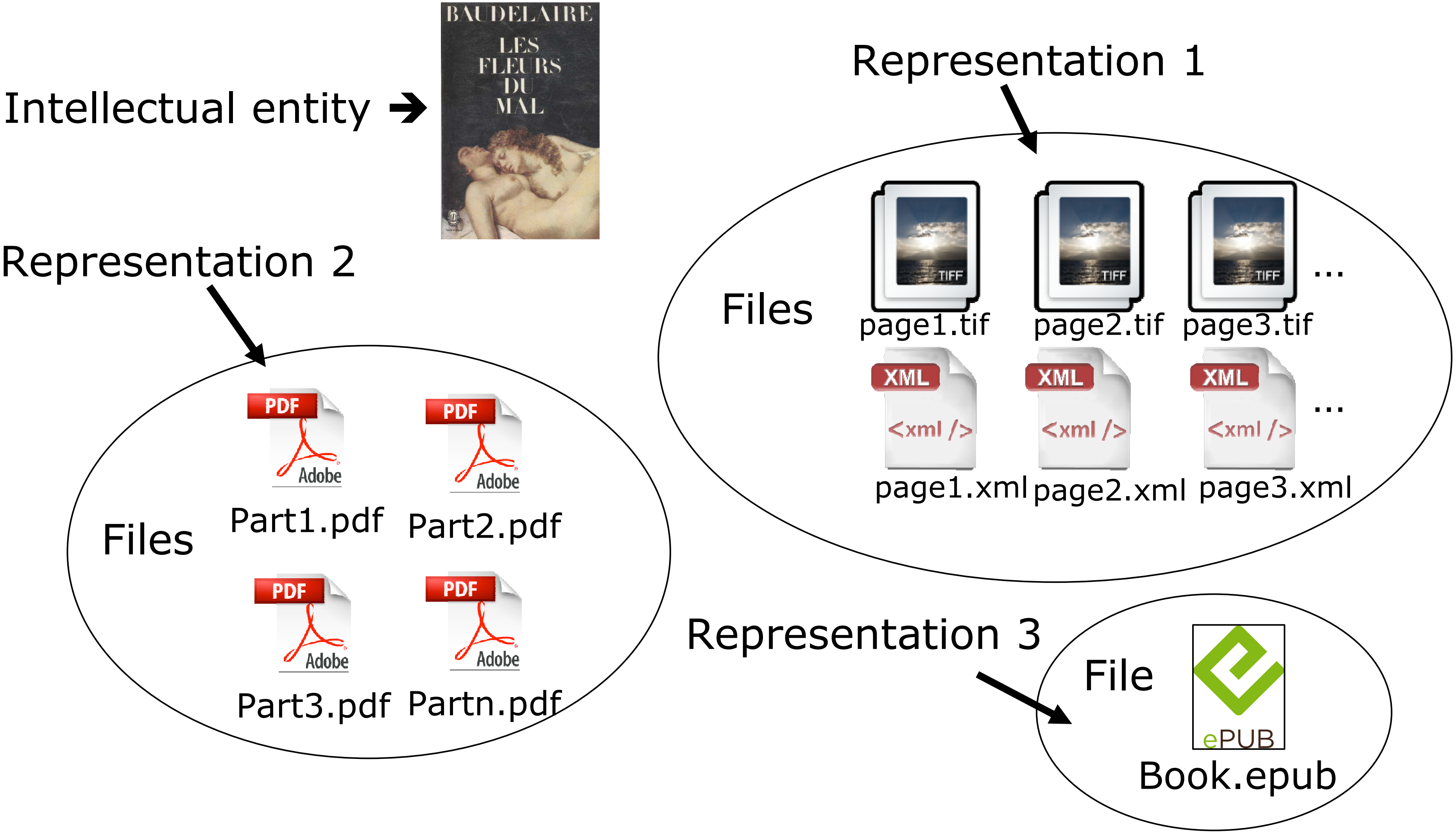
Examples:

- *Les Fleurs du Mal* by Charles Baudelaire (a book)
- "Maggie at the beach" (a photograph)
- The Library of Congress Website (a website)

- 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

THIS WILL CHANGE IN 3.0

Example: one content, 3 digital representations



Intellectual entity ➔

Representation 1

Files

page1.tif page2.tif page3.tif …

page1.xml page2.xml page3.xml …

Representation 2

Files

Part1.pdf Part2.pdf

Part3.pdf Partn.pdf

Representation 3

File

Book.epub

Object: high level semantic units

what technical information on it?

objectCharacteristics

where is it stored?

on which media?

storage

which object is it?

objectIdentifier

ark:/12148/btp6k102002g/f1

what kind of object?

objectCategory

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

significantProperties

what is my preservation strategy for this object?

preservationLevel

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

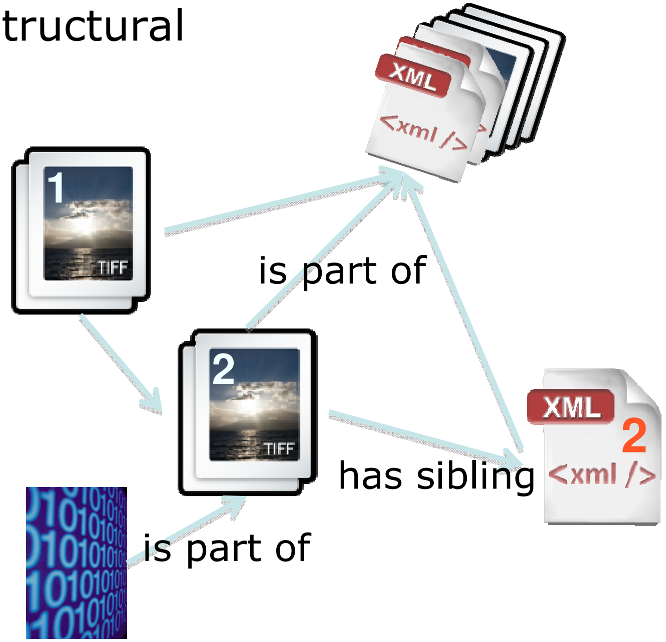
environment

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

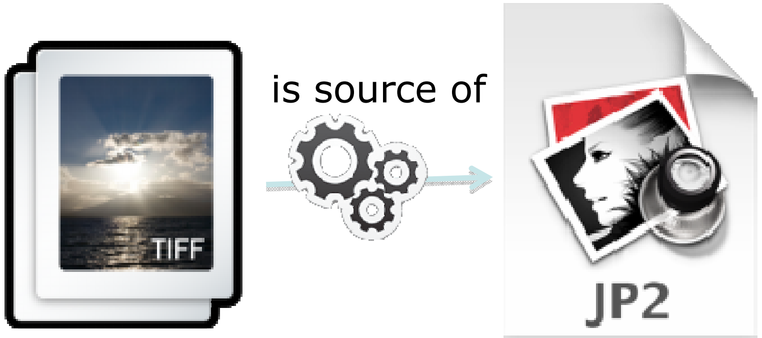


is part of

has sibling

is part of

- derivation



is source of

relationship

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

objectCharacteristics [for file or bitstream]

what checksum?

fixity

0a7d048211f3c4dc
e3a85c9c89a65651

what's its size
in bytes?

size

15484580

what format?

format

TXT PDF Adobe WAV MOV

what application was
used to create it?

creatingApplication

access restrictions on
this object?

(password, encryption…)

inhibitors

do I need to express
format specific information?

objectCharacteristicsExtension

textMD MIX MPEG7 …

is the object
directly renderable?

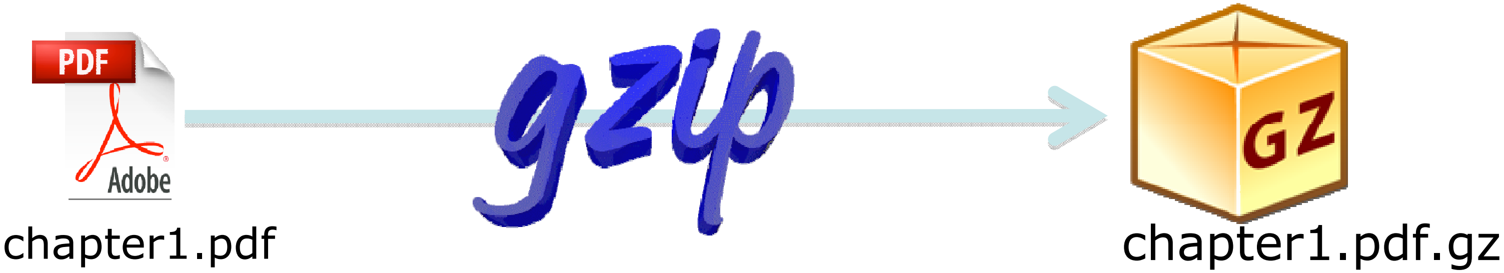
compositionLevel

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



chapter1.pdf

chapter1.pdf.gz

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

- compositionLevel = 1
- format = gzip
- size = 324,876 bytes
- messageDigest = [something else]

= different compositionLevels

Number of operations needed to access the primary data object

chapter1.pdf

composition Level			0
fixity	Message Digest Algorithm		SHA-1
fixity	Message Digest		[big string]
Fixity	Message Digest Originator		Submitter
Size			500000
format	format Designation	format Name	PDF
format	format Designation	format Version	1.2

chapter1.pdf.gz

composition Level			1
fixity	message Digest Algorithm		SHA-1
fixity	message Digest		[another string]
fixity	message Digest Originator		Repository
size			324876
format	format Designation	format Name	gzip
format	format Designation	format Version	1.2.3

format

Features:

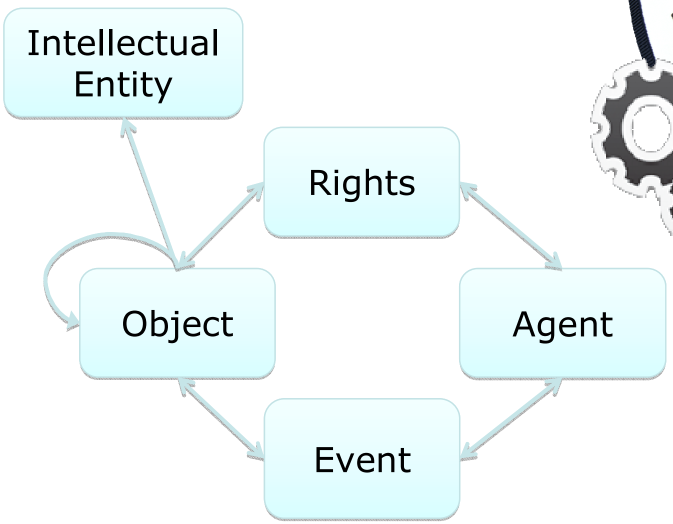
1. **Basic information** about the format
2. Link to some more detailed description in a **format registry**

semantic units	sample description
format	
formatDesignation (O,NR)	
formatName (M,NR)	`image/tiff`
formatVersion (O,NR)	`6.0`
formatRegistry	
formatRegistryName (M,NR)	`PRONOM`
formatRegistryKey (M,NR)	`fmt/353`
formatRegistryRole (O,NR)	`format specifications`
formatNote (O,R)	`http://www.nationalarchives.gov.uk`

objectCharacteristicsExtension: an example

```
<premis:mdSec>
 <premis:mdWrap MDTYPE="TEXTMD" MIMETYPE="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:markup_language>http://www.loc.gov/standards/alto/ns-
v2</textmd:markup_language>
   </textmd:textMD>
  </premis:xmlData>
 </premis:mdWrap>
</premis:mdSec>
```

Event examples



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?)

- 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 Objects 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

Event: high level semantic units

eventIdentifier (M,NR)

eventType (M,NR)

eventDateTime (M,NR)

eventDetail (O,NR)

eventOutcomeInformation (O,R)

linkingAgentIdentifier (O,R)

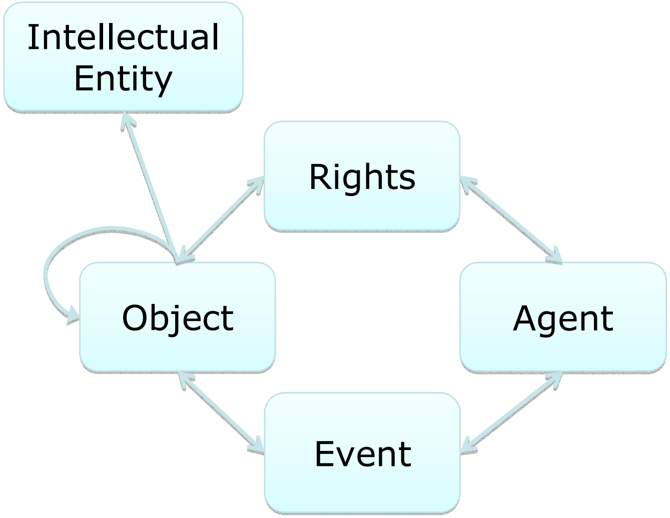
linkingObjectIdentifier (O,R)

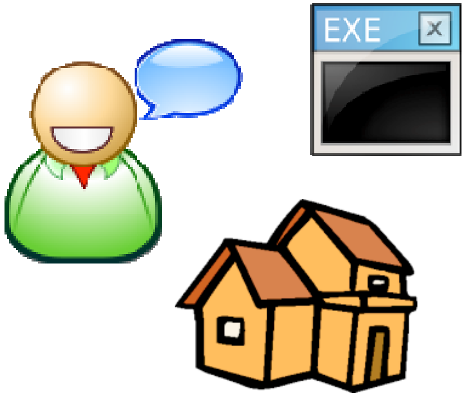
eventOutcomeInformation

eventOutcomeInformation	This event has an outcome.
eventOutcome	it has processed sucessfully.
eventOutcomeDetail	but how precisely?
eventOutcomeDetailNote	here is the machine response in plain text.
eventOutcomeDetail Extension	or here is the response in structured fashion

eventOutcomeInformation	**Sample description** validation event
eventOutcomeInformation	
eventOutcome	`validation process successful`
eventOutcomeDetail	
eventOutcomeDetailNote	`well-formed and valid`
	`(or)`
eventOutcomeDetail Extension	`<Whole XML output of JHOVE>`

Agent examples





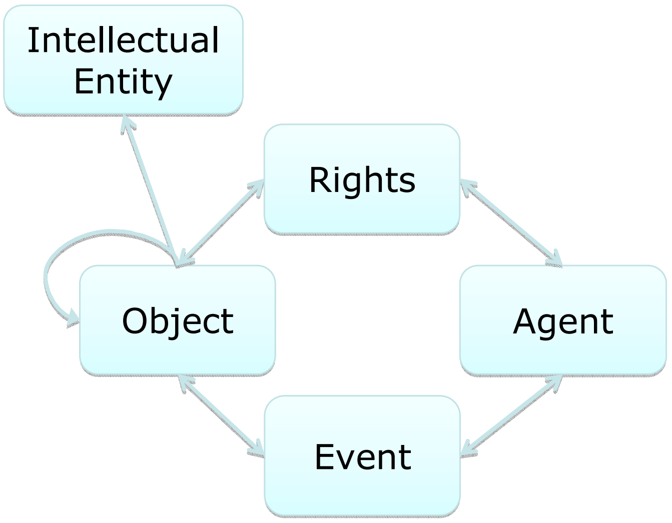
Examples:

- 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	Sample description
agentIdentifier	
agentIdentifierType	URI
agentIdentifierValue	info:bnf/spar/agent/jhove_1_5
agentName	JHOVE 1.5
agentType	software
agentNote	Release notes: http://sourceforge.net/projects/jhove/files/jhove/JHOVE%201.5/RELEASENOTES
agentExtension	

Rights statement examples



Intellectual Entity

Rights

Object

Agent

Event

- 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

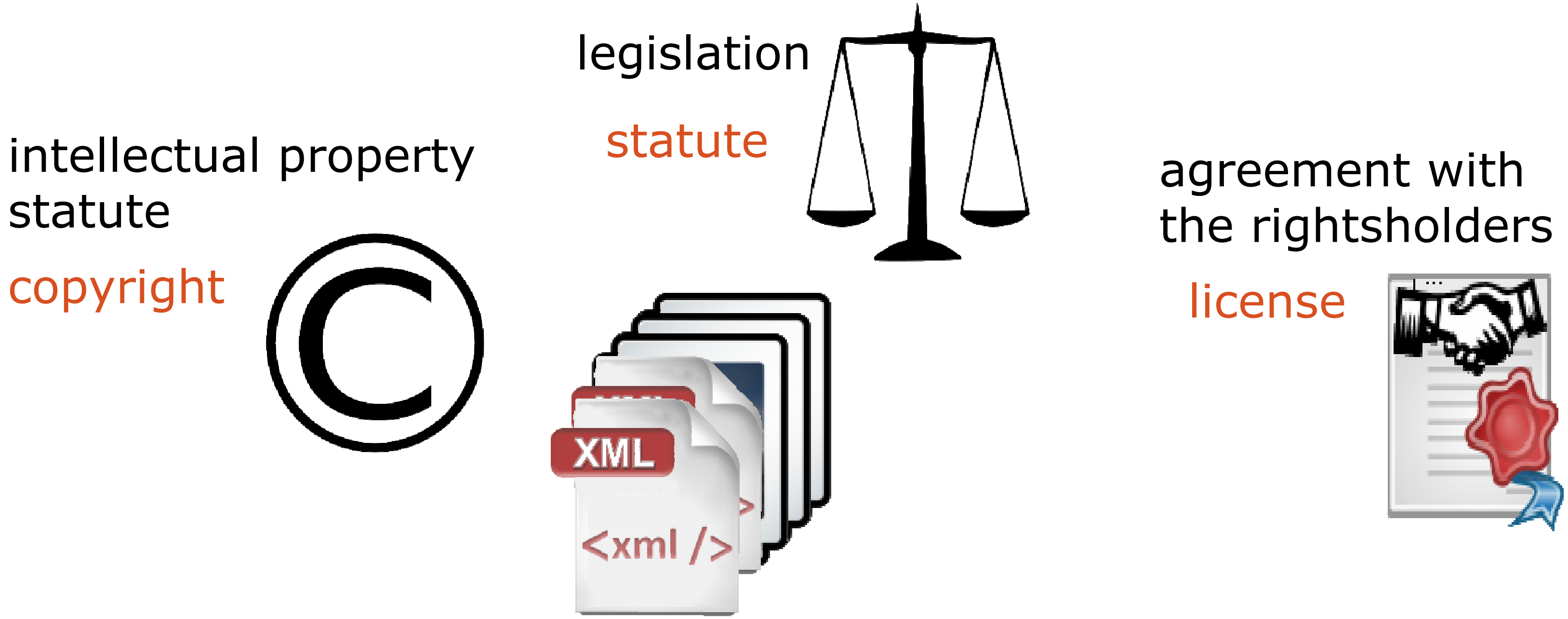
rightsStatement

- rightsStatementIdentifier
- rightsBasis
- copyrightInformation
- licenseInformation
- statuteInformation
- rightsGranted
- linkingObjectIdentifier
- linkingAgentIdentifier

rightsExtension

Either rightsStatement or rightsExtension must be present

rightsStatement: 3 possible rights bases



legislation

statute

intellectual property statute

copyright

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	what action is allowed?
restriction	on which conditions?
termOfGrant	
startDate	from when to when?
endDate	

rightsGranted	Sample description
rightsGranted	
act	dissemination
restriction	rightsholder must be notified
termOfGrant	
startDate	2010-05-05
endDate	2015-05-04

Sample data dictionary entry

Is it a container unit?

What does it contain?

Why should it be recorded?

How should it be recorded? constraints and examples

How should it be **provided**?

Some implementation guidelines

Semantic unit	size		
Semantic components	None		
Definition	The size in bytes of the file or bitstream stored in the repository.		
Rationale	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.		
Data constraint	Integer		
Object category	**Representation**	**File**	**Bitstream**
Applicability	Not applicable	Applicable	Applicable
Examples		2038927	
Repeatability		Not repeatable	Not repeatable
Obligation		Optional	Optional
Creation/ Maintenance notes	Automatically obtained by the repository.		
Usage notes	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.		

What’s next?

PREMIS OWL ontology
PREMIS 3.0 evolutions

PREMIS OWL ontology in a nutshell

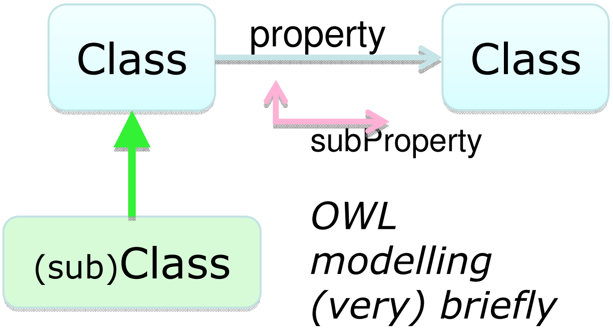
- Purpose
  - Providing the community with an RDF serialization of the PREMIS data model and dictionary
  - While remaining as close as possible to the data dictionary's clearly defined semantics

RDF modelling in 3 words:

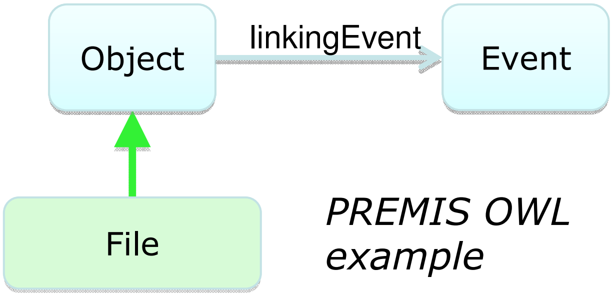
- Everything modelled under the form of

  subject-verb-object
- But what objects? what verbs? what objects?

  → role of vocabularies & ontologies

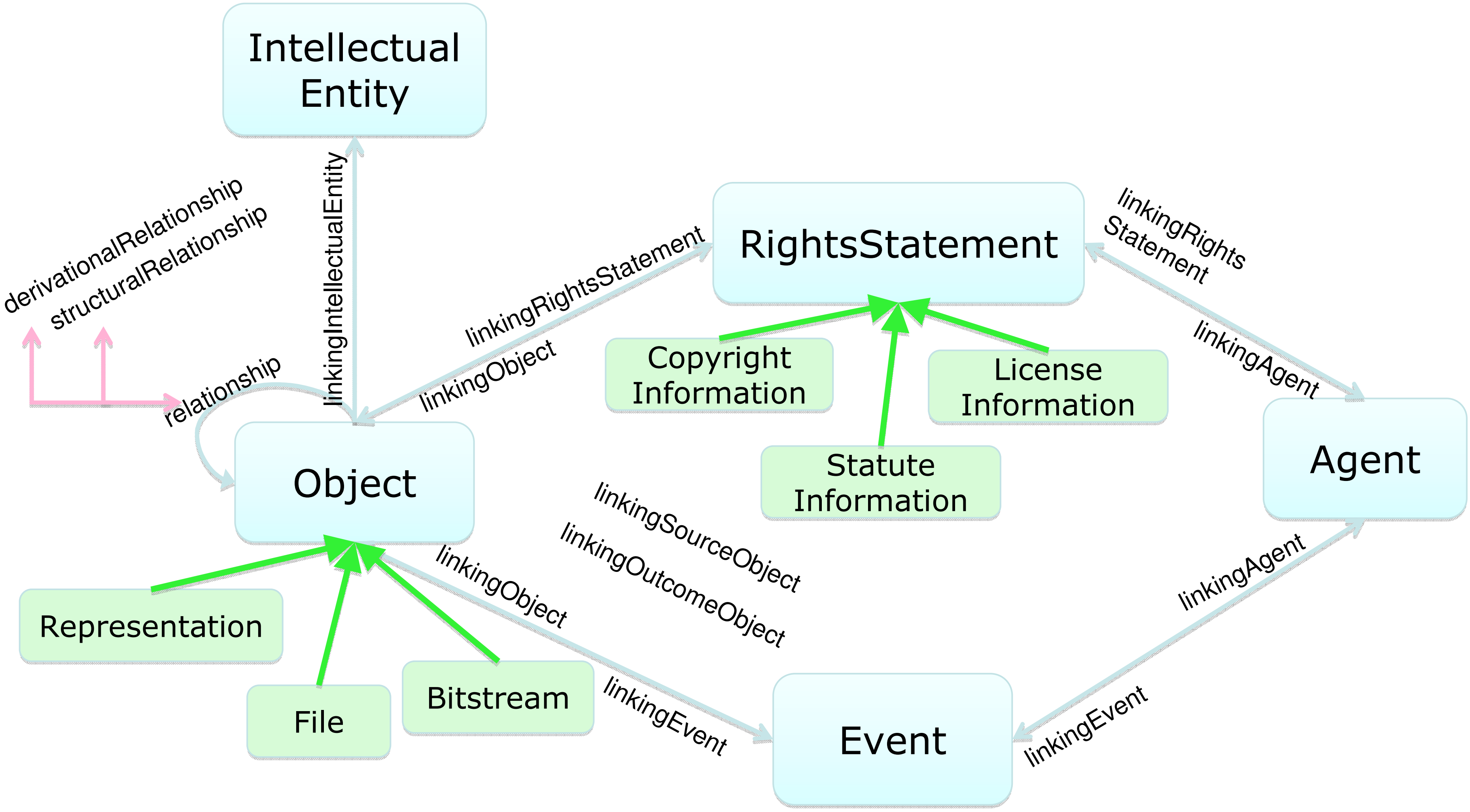


OWL modelling (very) briefly



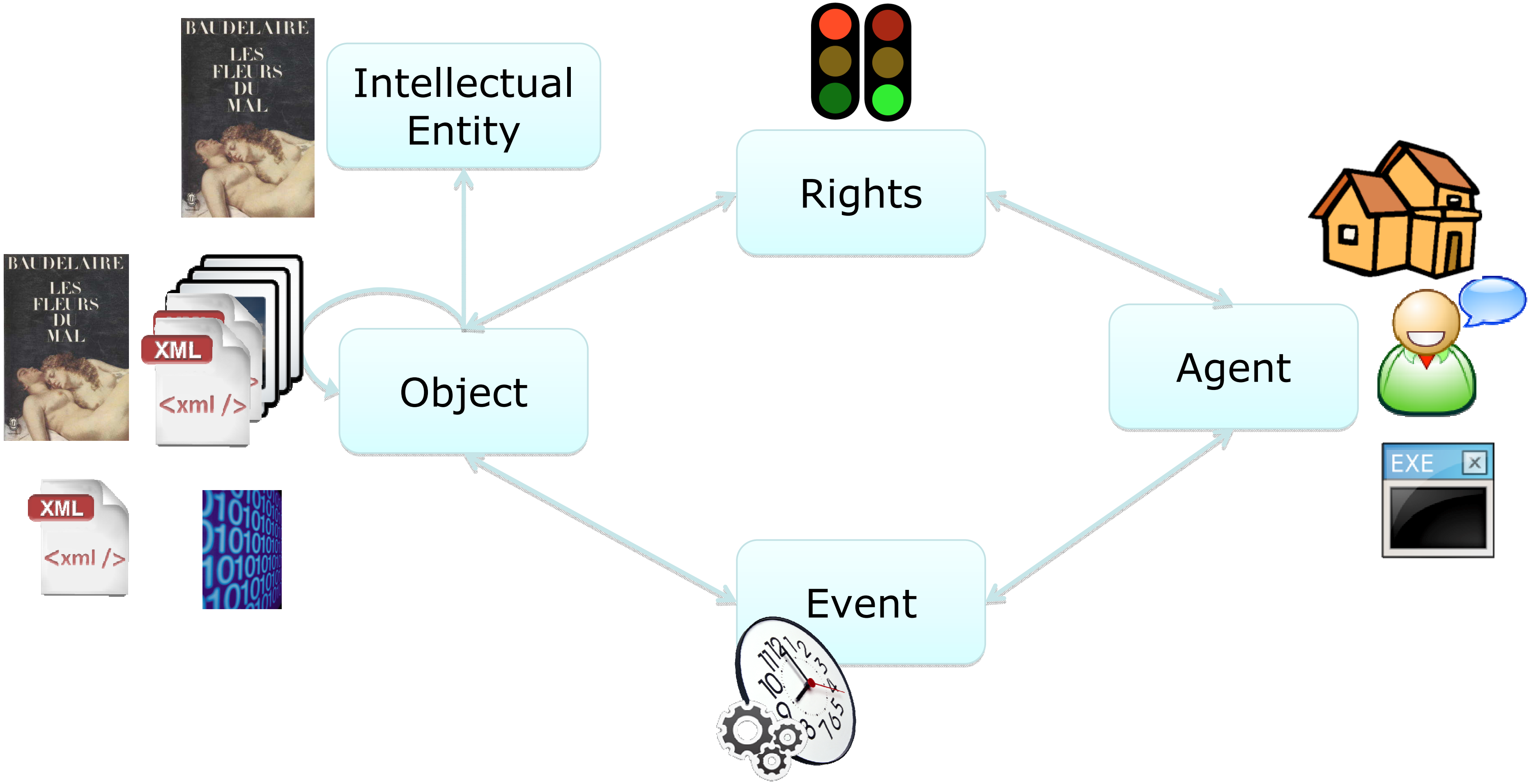
PREMIS OWL example

PREMIS ontology: key decisions



PREMIS 3.0: evolution of the data model

Intellectual entities become a category of object



Intellectual Entity

Rights

Object

Agent

Event

PREMIS 3.0: rights changes (work in progress)

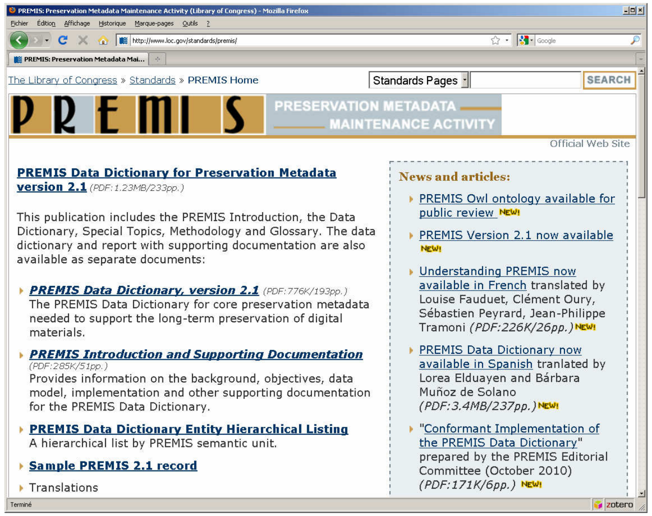
- rightsStatement
- rightsBasis
- copyrightInformation
    - copyrightDocumentationIdentifier
- licenseInformation
    - licenseDocumentationIdentifier
- statuteInformation
    - statuteDocumentationIdentifier
- otherRightsInformation
    - otherRightsBasis
    - otherRightsApplicableDates
- rightsGranted
    - act
    - restriction
    - termOfGrant
        - startDate
        - endDate
    - termOfRestriction
        - startDate
        - endDate

- 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


PREMIS: Preservation Metadata Maintenance Activity (Library of Congress) - Mozilla Firefox
Fichier
Édition
Affichage
Historique
Marque-pages
Outils
?
http://www.loc.gov/standards/premis/
Google
PREMIS: Preservation Metadata Mai...
Standards Pages
SEARCH
Terminé
zotero

The Library of Congress » Standards » PREMIS Home

PREMIS PRESERVATION METADATA MAINTENANCE ACTIVITY

Official Web Site

PREMIS Data Dictionary for Preservation Metadata version 2.1 *(PDF:1.23MB/233pp.)*

This publication includes the PREMIS Introduction, the Data Dictionary, Special Topics, Methodology and Glossary. The data dictionary and report with supporting documentation are also available as separate documents:

- ***PREMIS Data Dictionary, version 2.1*** *(PDF:776K/193pp.)* The PREMIS Data Dictionary for core preservation metadata needed to support the long-term preservation of digital materials.
- ***PREMIS Introduction and Supporting Documentation*** *(PDF:285K/51pp.)* Provides information on the background, objectives, data model, implementation and other supporting documentation for the PREMIS Data Dictionary.
- **PREMIS Data Dictionary Entity Hierarchical Listing** A hierarchical list by PREMIS semantic unit.
- **Sample PREMIS 2.1 record**
- Translations

News and articles:

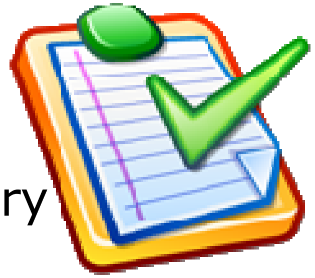
- PREMIS Owl ontology available for public review NEW!
- PREMIS Version 2.1 now available NEW!
- Understanding PREMIS now available in French translated by Louise Fauduet, Clément Oury, Sébastien Peyrard, Jean-Philippe Tramoni *(PDF:226K/26pp.)* NEW!
- PREMIS Data Dictionary now available in Spanish tranlated by Lorea Elduayen and Bárbara Muñoz de Solano *(PDF:3.4MB/237pp.)* NEW!
- "Conformant Implementation of the PREMIS Data Dictionary" prepared by the PREMIS Editorial Committee (October 2010) *(PDF:171K/6pp.)* NEW!

- 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
  http://www.loc.gov/standards/premis/premis-conformance-oct2010.pdf
- 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):

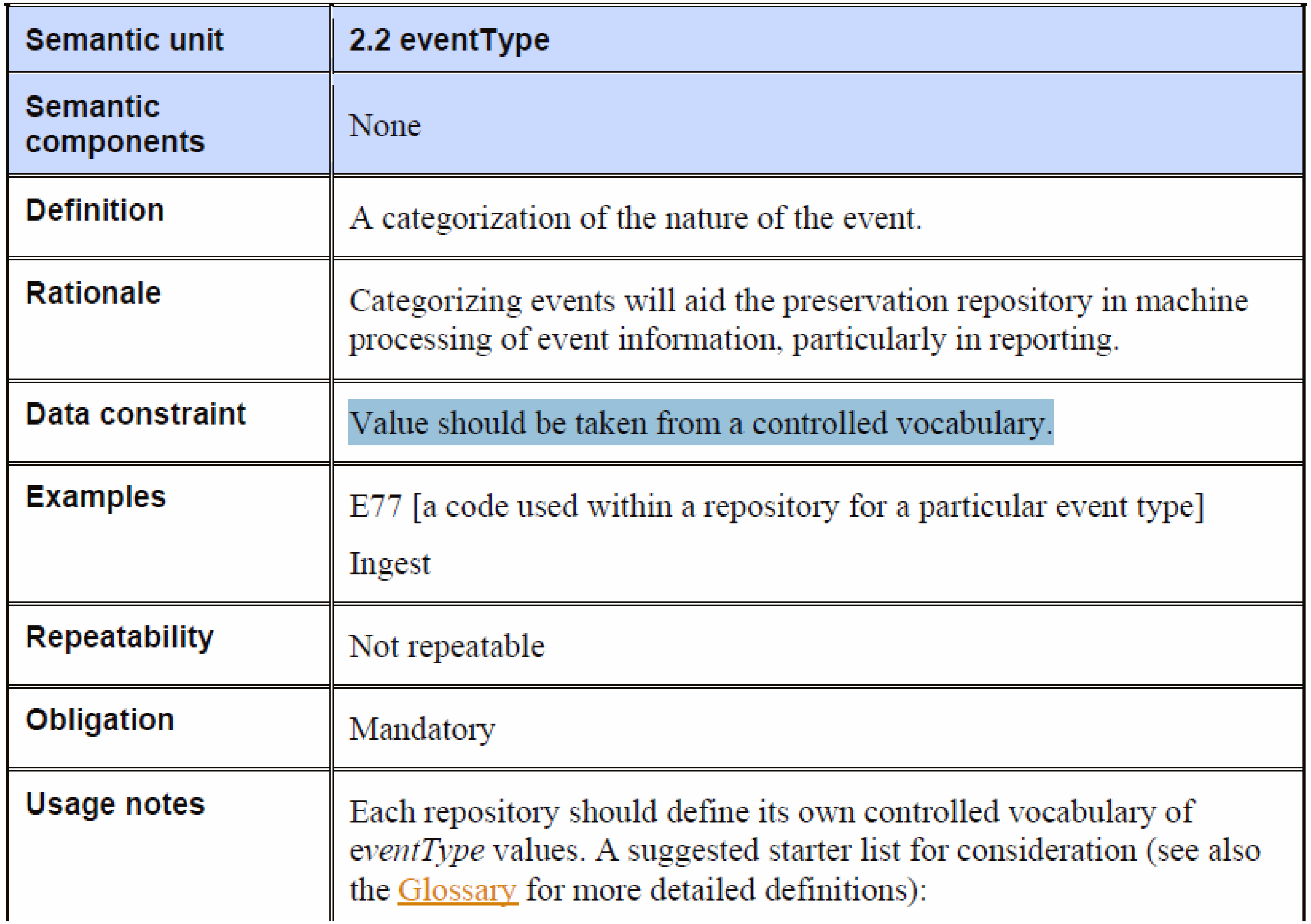
    import = the repository can manage PREMIS conformant information

    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



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

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
- 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.peyard@bnf.fr