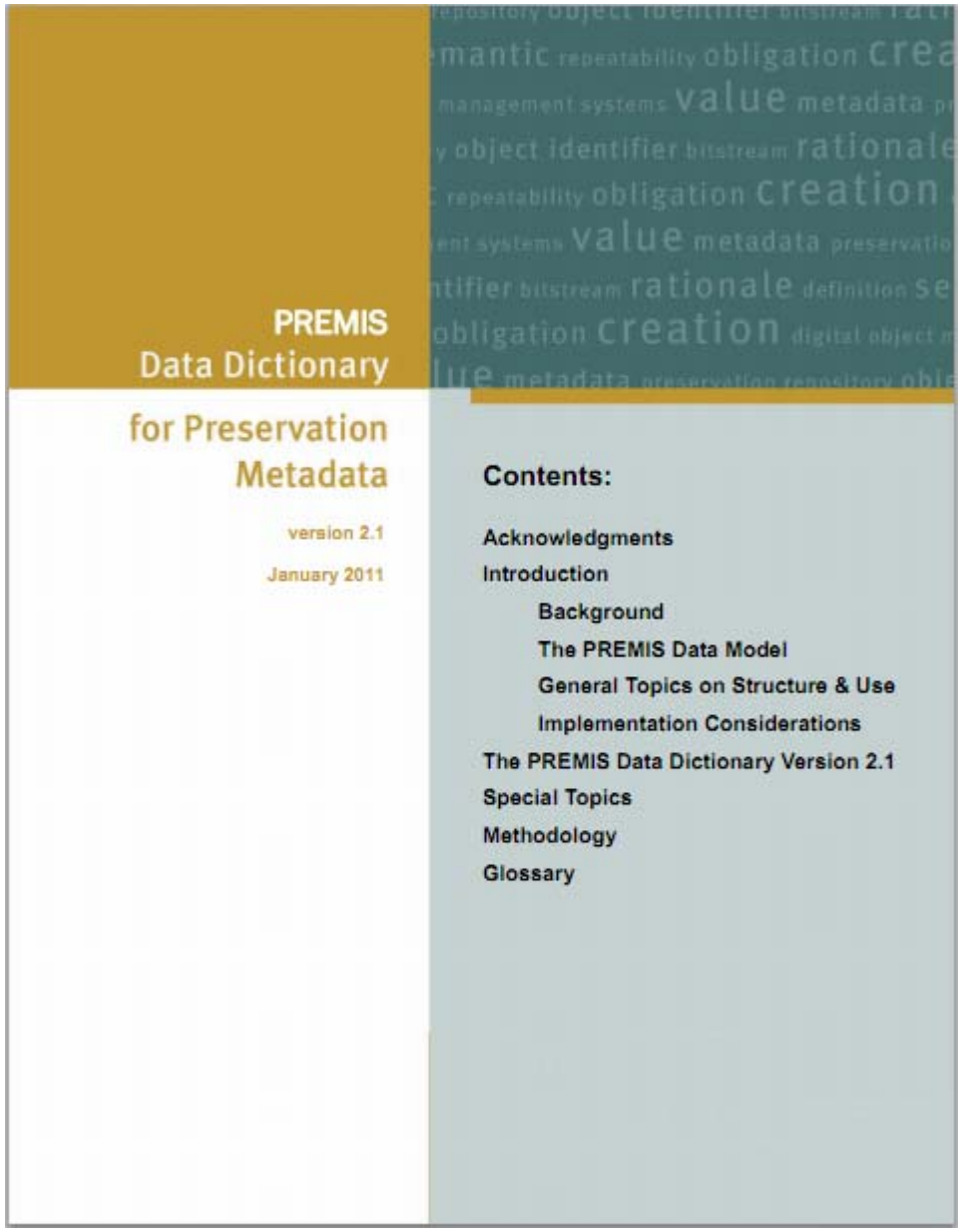


PREMIS

PREservation Metadata Implementation Strategies

WORKSHOP

db Metadaten & Vokabularien

24.-25. November 2011, Graz

Universitätszentrum Wall

Metadaten & Vokabularien 24.-25. November 2011, Graz

Thanks to the Organizers and Sponsors
for providing support to this PREMIS tutorial

Presenter
Angela Di Iorio

SAPIENZA UNIVERSITÀ DI ROMA *Digital Library*

http://sapienzadigitallibrary.uniroma1.it/

angeladiiorio@gmail.com
angela.diiorio@uniroma1.it

some of content's materials were excerpted from
"PREMIS Tutorials and Implementation Panel Presentations for Public Use"
http://www.loc.gov/standards/premis/tutorials.html

…and properly contextualized for the objectives and duration of this workshop

SUMMARY

- Background and context
- Data model and internal connections
- Data dictionary and defined entities
- Implementation approaches and issues
- Working with PREMIS
- PREMIS evolution

Background and context of PREMIS

The LONG TERM DIGITAL PRESERVATION connection:

- The existence of a digital record depends on its correct preservation
- The correct digital preservation starts from the born of digital resource
- The more the awareness of digital community is growing around the preservation, and the more the implementation initiatives had highlighted the need of a standardized metadata framework

PREMIS PREservation Metadata Implementation Strategies

Background and context of PREMIS

Pre-2002: various preservation metadata element sets released

- Different scopes, purposes, underlying models/assumptions
- No international standard; little consolidation of expertise/best practice

June 2002: Preservation Metadata Framework

- International working group (jointly sponsored by OCLC, RLG)
- Comprehensive, high-level description of types of preservation metadata
- Used OAIS reference model as starting point
- Set of "prototype" preservation metadata elements
- Consensus-based foundation for developing formal preservation metadata specifications

… but not an "off-the-shelf, ready to implement" solution

Post-2002: Need implementable preservation metadata, with guidelines for application and use, relevant to a wide range of digital preservation systems and contexts

- Motivated formation of the **PREMIS Working Group**

PREMIS Working group

- June 2003: OCLC, RLG sponsored new international working group:
  - ***PREMIS: Preservation Metadata: Implementation Strategies***

- Membership:
  - **more than 30 experts from 5 countries, representing libraries, museums, archives, government agencies, and the private sector**
  - Co-Chairs: Priscilla Caplan (FCLA), Rebecca Guenther (LC)

- Objective 1: Identify and evaluate alternative strategies for encoding, storing, managing, and exchanging preservation metadata
  - PREMIS Survey Report (September 2004)

- Objective 2: Define implementable, core preservation metadata, with guidelines/recommendations for management and use

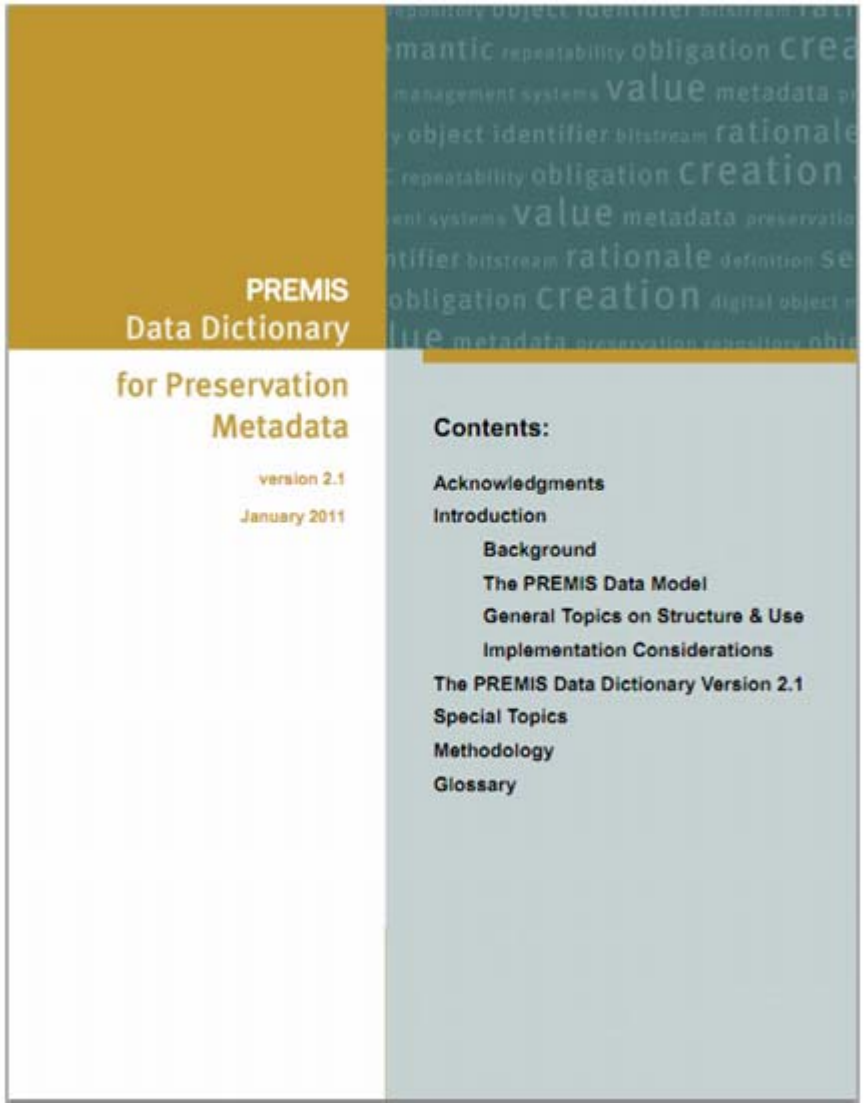
PREMIS PREservation Metadata Implementation Strategies

PREMIS Data Dictionary

May 2005: *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*

March 2008: *PREMIS Data Dictionary for Preservation Metadata, version 2.0*

January 2011: *PREMIS Data Dictionary for Preservation Metadata, version 2.1*

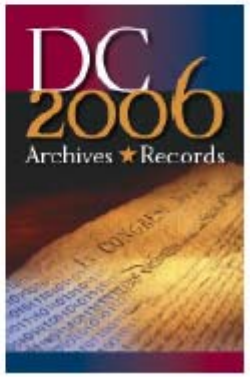


http://www.loc.gov/standards/premis/v2/premis-2-1.pdf

PREMIS acknowledgements

2005 British Conservation Awards: Digital Preservation Award

2006 Society of American Archivists Preservation Publication Award



PREMIS PREservation Metadata Implementation Strategies

PREMIS Data Dictionary

The initial conceiving of the Data Dictionary it was to define a set of preservation metadata that:

- *Supports the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context;*
- *Represents the information that most preservation repositories need to know to preserve digital materials over the long-term;*
- *Emphasizes "implementable metadata": rigorously defined, supported by guidelines for creation, management, and use, and oriented toward automated workflows; and*
- *Embodies technical neutrality: no assumptions made about preservation technologies, strategies, metadata storage and management, etc.*

Data Dictionary for Preservation Metadata: PREMIS version 2.1 [PREMISDD-2.1] : p. 1

The PREMIS Data Dictionary defines "preservation metadata" as

the information a repository
uses to support the digital preservation process.

[PREMISDD-2.1] : p. 3

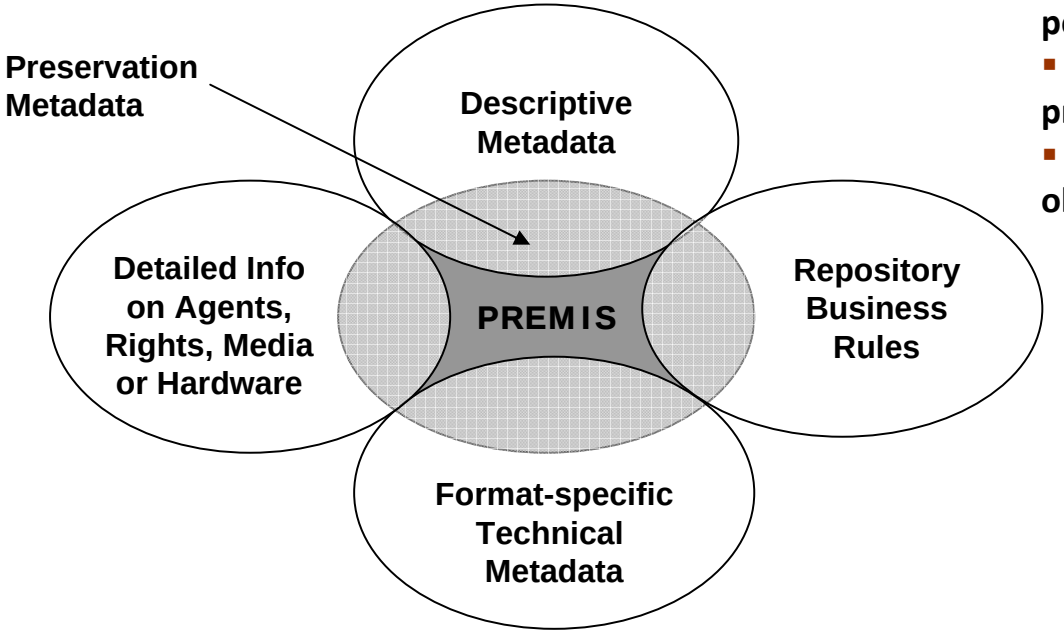
PREMIS PREservation Metadata Implementation Strategies

PREMIS core preservation metadata

"Implementable, core preservation metadata"

things that most working preservation repositories are likely to need to know in order to support digital preservation

[PREMISDD-2.1] : p. 3



- **Expands a number of the categories typically used to differentiate types of metadata: administrative (including rights and permissions), technical, and structural.**
- **Documents as much as possible the digital provenance (the history of an object)**
- **Documents the relationships among different objects within the preservation repository**

Understanding PREMIS by Priscilla Caplan for the Library of Congress, http://www.loc.gov/standards/premis/understanding-premis.pdf pp. 4-5

PREMIS PREservation Metadata Implementation Strategies

PREMIS Data Dictionary

"Implementable, core preservation metadata"

things that most working preservation repositories are likely to need to know in order to support digital preservation

CORE

- absolutely required under any circumstances;
- applicable to any type of repository implementing any type of preservation strategy;
- does not necessarily mean mandatory

IMPLEMENTABLE

- rigorously defined;
- supported by usage guidelines/recommendations;
- emphasis on automated workflows (automatically supplied and processed by the Repository)
- coded values from an authority list are preferred over textual descriptions
- multi-level technical neutrality: no assuptions about system, data management, preservation strategies

Because of the emphasis on the need to know rather than the need to record or represent in any particular way, the group preferred to use the term "semantic unit" rather than "metadata element." The Data Dictionary names and describes semantic units.

[PREMISDD-2.1] : p. 4

PREMIS PREservation Metadata Implementation Strategies

PREMIS Data Dictionary

What PREMIS DD is

- Common data model for organizing/thinking about preservation metadata
- Guidance for local implementations
- Standard for exchanging information packages between repositories
- Reference for core preservation metadata

What PREMIS DD is not

- Out-of-the-box solution: need to instantiate metadata elements in 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 repository
- Rights management: limited to permissions regarding actions taken within repository

PREMIS PREservation Metadata Implementation Strategies

PREMIS Data Dictionary

BENEFITS

- supplies a critical piece of the digital preservation infrastructure, and is a building block with which effective, sustainable digital preservation strategies can be implemented;

- is the first comprehensive technical specification for preservation metadata produced from an international, cross-domain, consensus-building process;

- is widely applicable across all sorts of institutions, digital preservation contexts, and system implementations;

- is oriented toward practical implementation;

- is supported by the PREMIS Maintenance Activity, which provides a central destination for PREMIS related information and resources, and hosts the PREMIS Implementers' Group discussion list.

PREMIS PREservation Metadata Implementation Strategies

PREMIS Maintenance Activity

Web site

- Permanent Web presence, hosted by Library of Congress
- Central destination for PREMIS-related info, announcements, resources
- Home of the PREMIS Implementers' Group (PIG) discussion list

PREMIS Editorial Committee

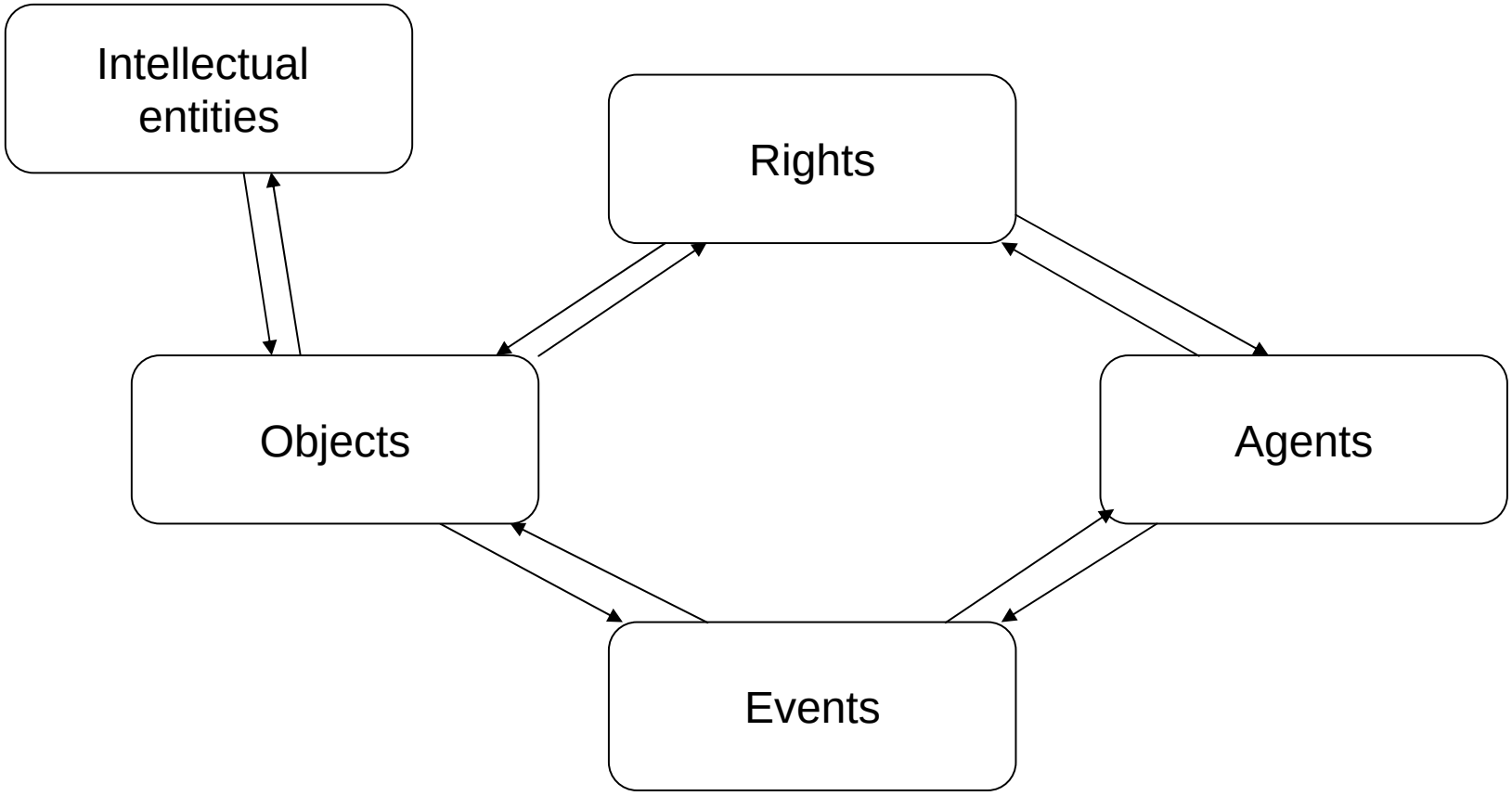
- Set directions/priorities for PREMIS development
- Considers proposals for changes
- Coordinates revisions of Data Dictionary and XML schema

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

PREMIS PREservation Metadata Implementation Strategies

PREMIS Data Model

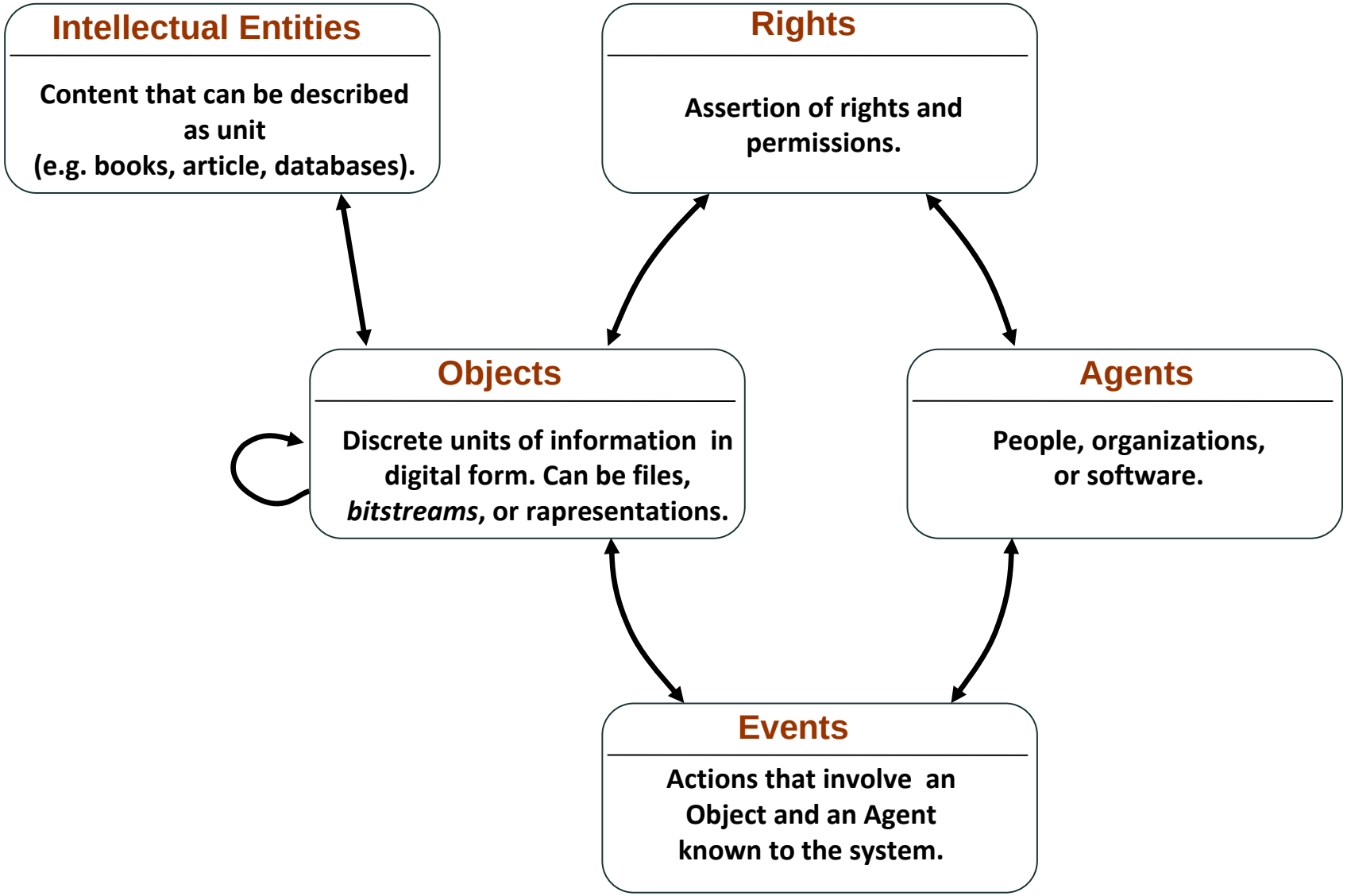
Entities, “things” relevant to digital preservation that are described by preservation metadata (Intellectual Entities, Objects, Events, Rights, Agents) [boxes]
Properties of Entities (semantic units)
Relationships between Entities [arrows]



PREMIS PREservation Metadata Implementation Strategies

PREMIS Data Model

Understanding PREMIS by Priscilla Caplan for the Library of Congress, http://www.loc.gov/standards/premis/understanding-premis.pdf p. 8



PREMIS PREservation Metadata Implementation Strategies

Intellectual entity

Intellectual Entities

Content that can be described as unit (e.g. books, article, databases).

Intellectual Entity: a set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database.
An Intellectual Entity can include other Intellectual Entities; for example, a Web site can include a Web page; a Web page can include an image.
An Intellectual Entity may have one or more digital representations.

Examples:

- *The Chamber* by John Grisham(an ebook)
- "Maggie at the beach" (a photograph)
- The Library of Congress Website (a website)

Not fully described in PREMIS DD, but can be linked to in metadata describing digital representation… this will change in forthcoming version 3.0

PREMIS PREservation Metadata Implementation Strategies

Object entity

Object (or Digital Object): a discrete unit of information in digital form

Objects

Discrete units of information in digital form. Can be files, *bitstreams*, or rapresentations.

Objects are what are actually stored and managed in the preservation repository.

Most of PREMIS is devoted to describing digital objects.

The information that can be recorded includes:

- a unique identifier for the object (type and value),
- fixity information such as a checksum (message digest) and the algorithm used to derive it,
- the size of the object,
- the format of the object, which can be specified directly or by linking to a format registry,
- the original name of the object,
- information about its creation,
- information about inhibitors,
- information about its significant properties,
- information about its environment (see below),
- where and on what medium it is stored,
- digital signature information,
- relationships with other objects and other types of entities.

PREMIS PREservation Metadata Implementation Strategies

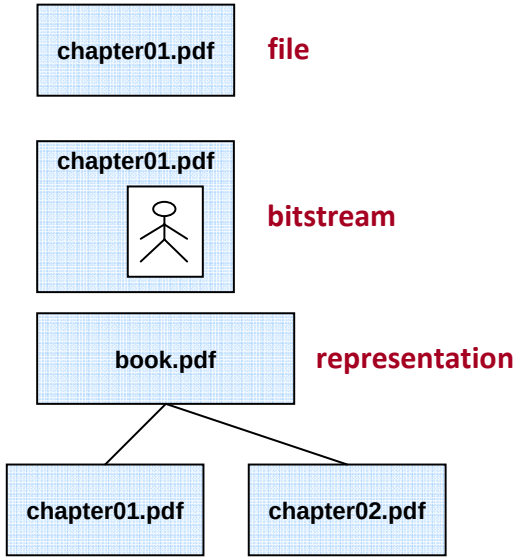
Object entity

Objects

Discrete units of information in digital form. Can be files, *bitstreams*, or rapresentations.

Examples:

- a PDF file
- a book composed of several XML files and many images
- TIFF file containing a header and 2 images
- a ZIP file

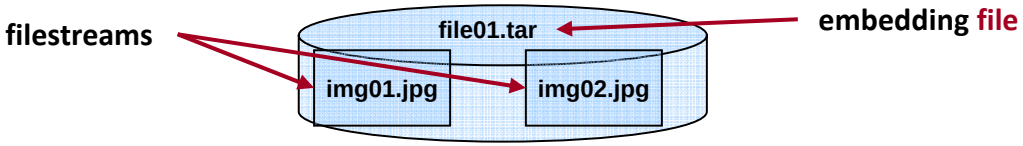


The Object entity has three subtypes: file, bitstream, and representation.

A **file** is a named and ordered sequence of bytes that is known by an operating system. A file can be zero or more bytes and has a file format, access permissions, and file system characteristics such as size and last modification date.

A **bitstream** is contiguous or non-contiguous data within a file that has 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.

A **representation** is the set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity. For example, a journal article may be complete in one PDF file; this single file constitutes the representation. Another journal article may consist of one SGML file and two image files; these three files constitute the representation. A third article may be represented by one TIFF image for each of 12 pages plus an XML file of structural metadata showing the order of the pages; these 13 files constitute the representation.



PREMIS PREservation Metadata Implementation Strategies

Agent entity

Agents

People, organizations, or software.

Agent: person, organization, or software program/system associated with Events in the life of an Object, or with Rights (permission statement) attached to an Object.

- Agents are associated only indirectly to Objects through Events or Rights

- Not defined in detail in PREMIS DD because is not considered core preservation metadata beyond identification

Examples:

- Angela Di Iorio (a person)
- Università di Roma la Sapienza (an organization)
- Sapienza Digital Library System (a system)
- BRI-DGE version 2.0 (a software program)

PREMIS PREservation Metadata Implementation Strategies

Event entity

Events

Actions that involve an Object and an Agent known to the system.

Event: 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

Examples:

- Validation Event: use JHOVE tool to verify that chapter01.pdf is a valid PDF file

- Ingest Event: transform an OAIS SIP into an AIP (one Event or multiple Events?)

- Migration Event: create a new version of an Object in an up-todate format

PREMIS PREservation Metadata Implementation Strategies

Rights entity

Rights

Assertion of rights and permissions.

Rights: assertions of one or more rights or permissions pertaining to an Object and/or Agent.

- 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 exclusively on permissions that take the form: Agent X grants Permission Y to the repository in regard to Object Z.

- Rights may be associated with copyright, license or contract

Example:

Dario Fo´ grants Sapienza Digital Library archival system permission, for preservation purposes, to make three copies of the archival records that are under his ownership.

PREMIS PREservation Metadata Implementation Strategies

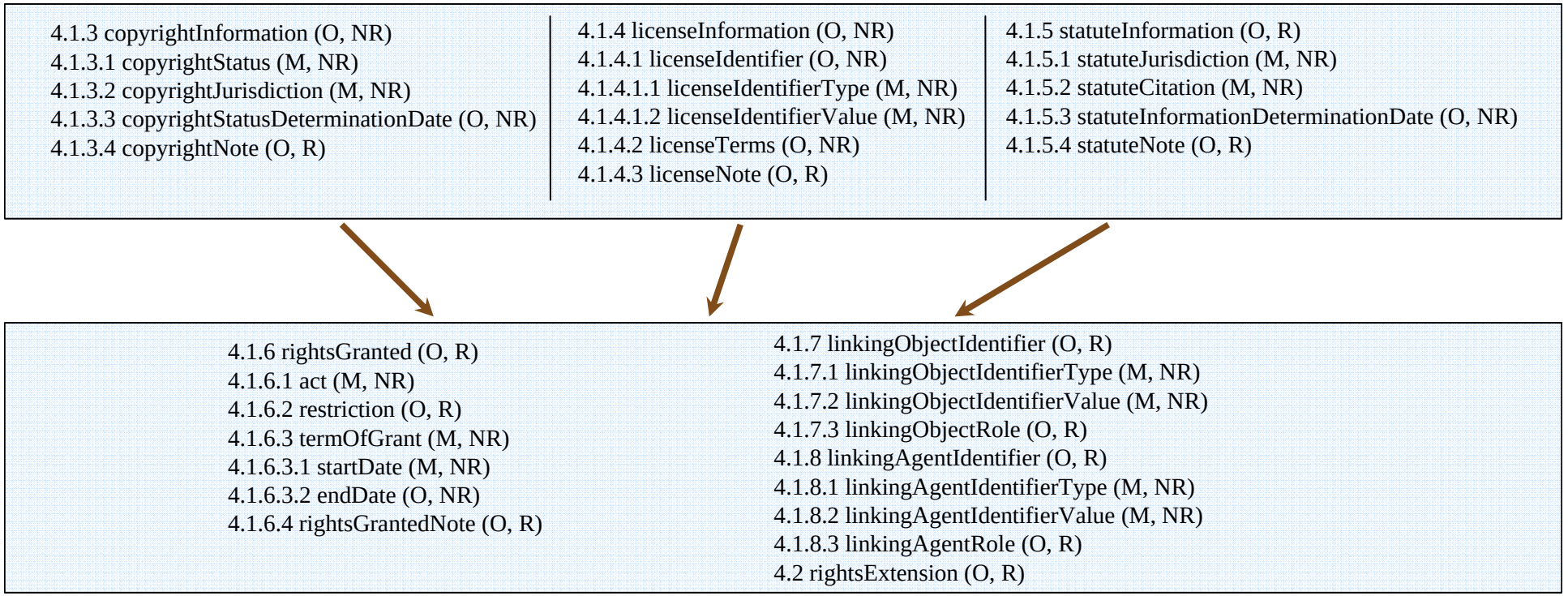
Rights entity

RightsBasis: copyright, statute, license

If case of... **copyrightInformation** container must be present

In case of... **licenseInformation** container must be present

In case of... **statuteInformation**n container must be present



Changes and integration to the Rights entity will be provided in forthcoming version 3.0

Semantic units

A semantic unit is a property of an Entity

- Something you need to know about an Object, Event, Agent, Right
- 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:

Extension containers

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

The extensibility mechanism

The set of semantic units where the extensibility is supported by the XML schemas is the following:

- significantProperties [Object entity]
- objectCharacteristics [Object entity]
- creatingApplication [within objectCharacteristics, Object entity]
- environment [within objectCharacteristics, Object entity]
- signatureInformation [Object entity]
- eventOutcomeDetail [within eventOutcomeInformation, Event entity]
- rights [Rights entity]
- agent [Agent entity] **(was added in version 2.1)**

The PREMIS use of Identifiers

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 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 PREservation Metadata Implementation Strategies

HowTo express Relationships

Relationships

Many different types of information relevant to preservation can be expressed as relationships:

- e.g., “A is part of B”, “A is scanned from B”, “A is a version of B”

PREMIS Data Dictionary supports expression of relationships between:

- Different Objects
- Across same level or different levels
- Structural: relationships between parts of a whole
- Derivation: relationships resulting from replication or transformation of an Object
- Different Entities

Relationships are established through reference to Identifiers of other Objects or Entities

Relationships between Objects

Relationships between Objects: Which, How, Why

WHICH Objects are related?

- relatedObjectIdentification: type, value
- relatedObjectSequence: documents “ordered” relationships: e.g., pages, chapters, slide #

HOW are the Objects related?

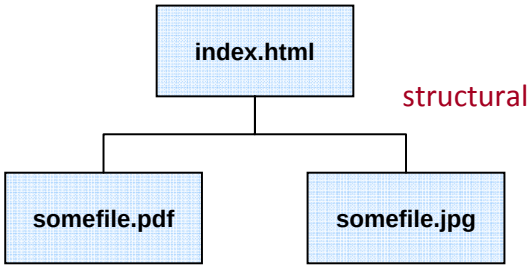
- relationshipType: structural, derivation
- relationshipSubType: “is part of”, “is source of”, “is derived from”

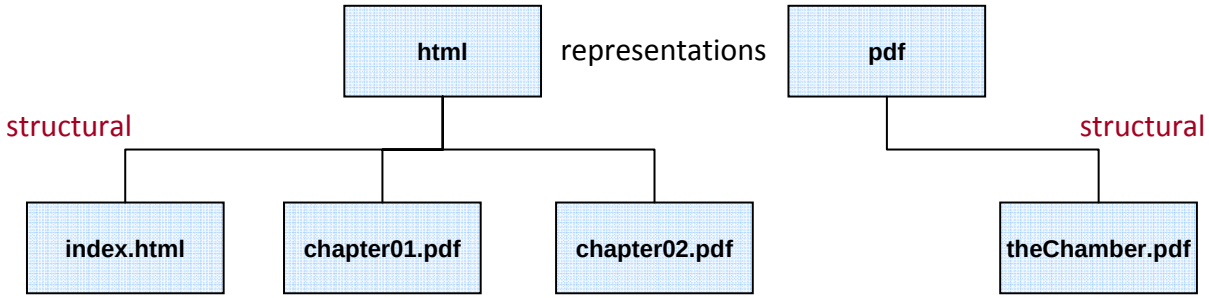
WHY are the Objects related?

- Was relationship result of an Event? (e.g., “migration”, “replication”)
- relatedEventIdentification: type, value
- relatedEventSequence: ordered sequence of Events
- Event 1: Convert Excel spreadsheet to ASCII tab-delimited file
- Event 2: Convert ASCII file to new spreadsheet format
- Avoids numerous bilateral format-to-format conversions

Relationships between Objects

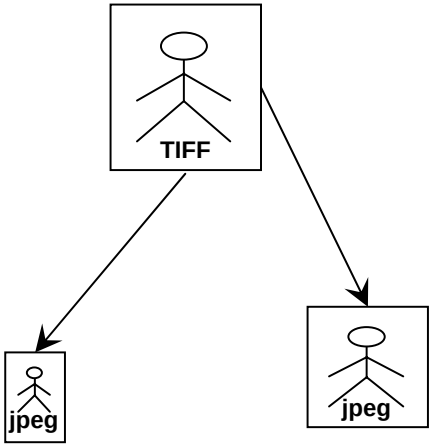
Structural relationships show relationships between parts of objects. The structural relationships between the files that constitute a representation of an Intellectual Entity are clearly essential preservation metadata.





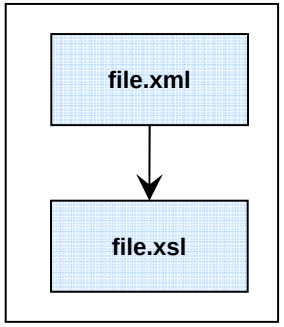
Relationships between Objects

Derivation relationships result from the replication or transformation of an Object. The intellectual content of the resulting Object is the same, but the Object's instantiation, and possibly its format, are different.
When file A of format X is migrated to create file B of format Y, a derivation relationship exists between A and B.



Relationships between Objects

A **dependency relationship** exists when one object requires another to support its function, delivery, or coherence of content. An object may require a font, style sheet, DTD, schema, or other file that is not formally part of the object itself but is necessary to render it. The Data Dictionary handles dependency relationships as part of the environment information, in the semantic units *dependency* and *swDependency.*
In this way requirements for hardware and software are brought together with requirements for dependent files to form a complete picture of the information or assets required for the rendering and/or understanding of the object.



Relationships between different Entities

Identifiers are used to link related Entities together:

- **linkingIntellectualEntityIdentifier**

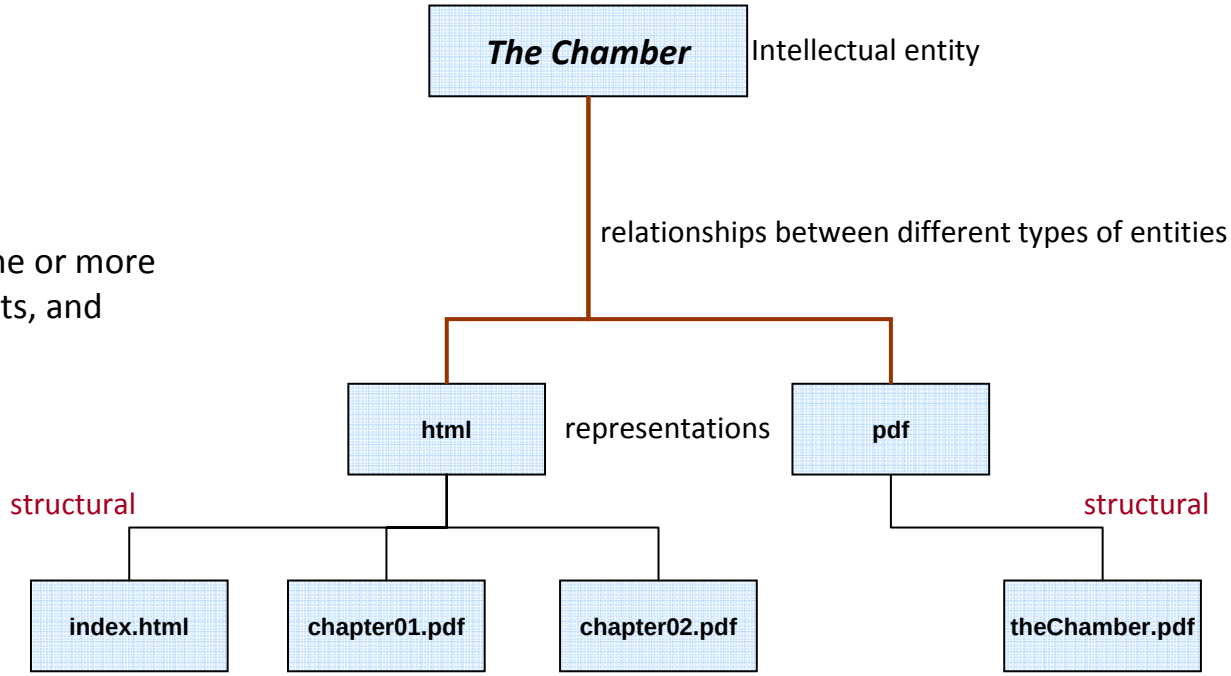
- **linkingRightsStatementIdentifier**

- **linkingEventIdentifier**

- **linkingAgentIdentifier**

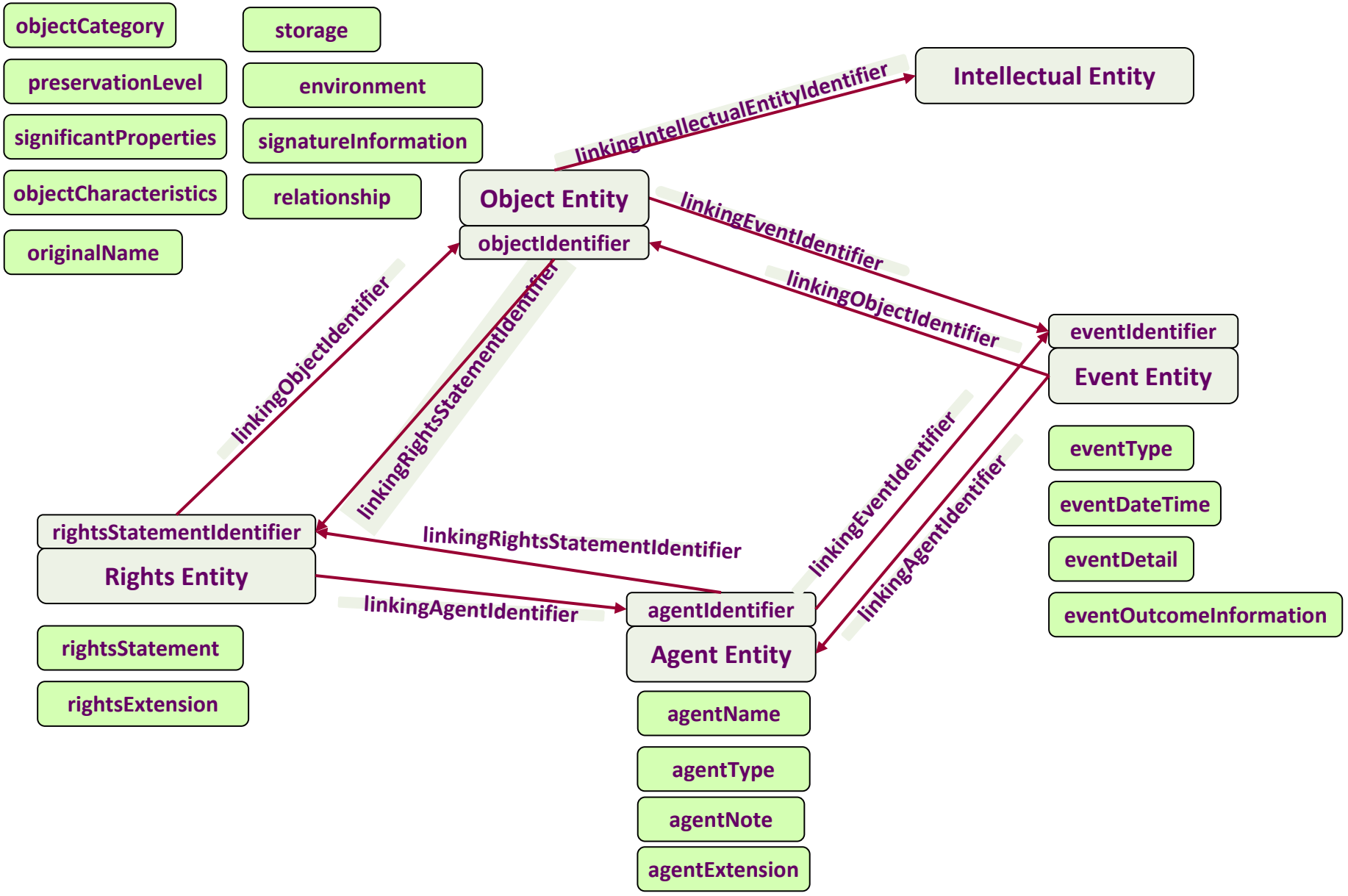
- **linkingObjectIdentifier**

For example, an Object can link to one or more Intellectual Entities, Rights statements, and Events via “linking” semantic units



PREMIS PREservation Metadata Implementation Strategies

Relationships between different Entities



Semantic units in Data Dictionary

Semantic unit	1.5 objectCharacteristics		
Semantic components	1.5.1 compositionLevel 1.5.2 fixity 1.5.3 size 1.5.4 format 1.5.5 creatingApplication 1.5.6 inhibitors 1.5.7 objectCharacteristicsExtension		
Definition	Technical properties of a file or bitstream that are applicable to all or most formats.		
Rationale	There are some important technical properties that apply to objects of any format. Detailed definition of format-specific properties is outside the scope of this Data Dictionary, although such properties may be included within *objectCharacteristicsExtension*.		
Data constraint	Container		
Object category	Representation	File	Bitstream
Applicability	Not applicable	Applicable	Applicable
Repeatability		Repeatable	Repeatable
Obligation		Mandatory	Mandatory
Usage notes	The semantic units included in *objectCharacteristics* should be treated as a set of information that pertains to a single object at a single *compositionLevel*. Object characteristics may be repeated		

Semantic units in Data Dictionary

Semantic unit	1.5.3 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		2038937	
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.		

PREMIS PREservation Metadata Implementation Strategies

Implementation example

Let´s have a look to the example:

http://www.loc.gov/standards/premis/louis-2-1.xml

PREMIS conformance

The importance of technical neutrality as a design principle for the Data Dictionary implies that any conformance requirements associated with the Dictionary will necessarily be lightweight.
But this is not to say that conformance is unimportant in a PREMIS context; in fact, there are a number of use cases where establishing shared expectations in regard to a PREMIS implementation is of practical benefit, including:

- Inter-repository data exchange
- Repository certification
- Shared registries
- Automation/reusable tools
- Vendor support

[PREMISDD-2.1] : pp. 15

PREMIS PREservation Metadata Implementation Strategies

PREMIS conformance

PRINCIPLES OF USE (SEMANTIC UNIT): A conformant implementation of a PREMIS semantic unit must follow all requirements and constraints prescribed in the *latest version of the Data Dictionary* for that semantic unit.
Specifically:

- If a metadata element shares the name of a PREMIS semantic unit, it must also share its definition. If a metadata element shares the definition of a PREMIS semantic unit but does not share its name, the repository must establish a mapping between the metadata element and its corresponding PREMIS semantic unit.
- Usage requirements specified in the Data Dictionary for a particular semantic unit must be observed. Repeatability, obligation (i.e., whether a semantic unit is mandatory), and applicability (bit stream, file, and representation) requirements can be made more stringent, but *not* more relaxed.

An implementation of a PREMIS semantic unit that fails to observe any of these principles is considered non-conformant.

The conformance statement is available at:
http://www.loc.gov/standards/premis/premis-conformance-oct2010.pdf

PREMIS PREservation Metadata Implementation Strategies

PREMIS conformance

PRINCIPLES OF USE (DATA DICTIONARY): A conformant implementation of the PREMIS Data Dictionary *at the minimum* must:

- Include the mandatory semantic units for any Data Model Entity (Objects, Events, Agents, or Rights) supported by the repository.

- Be able to recover all of the information specified in the mandatory PREMIS semantic units from the repository system (regardless of its specific implementation), and associate it with its corresponding Entity.

A repository's implementation of the PREMIS Data Dictionary that fails to observe any of these principles is considered non-conformant.

The conformance statement is available at:
http://www.loc.gov/standards/premis/premis-conformance-oct2010.pdf

PREMIS conformance internal/external

INTERNAL CONFORMANCE: PREMIS conformance as it relates to PREMIS-based information residing *within a repository*.
A repository that satisfies the Principles of Use at both the semantic unit and Data Dictionary levels is considered *internally conformant*.

EXTERNAL CONFORMANCE: PREMIS conformance as it relates to the exchange of PREMIS-based information *between repositories*. There are two forms of external conformance: *import* and *export*.

Import: A repository that is *import conformant* must be able to accept PREMIS-conformant information in the form provided by another repository, parse it, and allocate the information to its corresponding metadata elements in the local repository system, as well as associate it with the appropriate Entities.

Export: A repository that is *export conformant* must be able to extract PREMIS-conformant information from its local system, and provide it to another repository in an agreed-upon form, and associate it with its appropriate Entity.

PREMIS conformance degrees of freedom

Naming: A repository is free to implement PREMIS semantic units using names different from those defined in the Data Dictionary. (However, remember that *if* a metadata element does share the name of a PREMIS semantic unit, it must share its definition; see Principles of Use (Semantic Unit) above.)

Granularity: A repository is free to implement PREMIS semantic units at higher or lower levels of granularity than what is defined in the Data Dictionary. Put another way, a metadata element implemented by a repository can incorporate information from more than one PREMIS semantic unit, or alternatively, encompass only part of the information defined in a PREMIS semantic unit (e.g., if the information from a PREMIS semantic unit is distributed over multiple metadata elements).

Level of Detail: A repository is free to record more detailed information for a PREMIS semantic unit than what is defined in the Data Dictionary (although the information defined in the Data Dictionary should be a subset of the more detailed information recorded by the repository).

PREMIS PREservation Metadata Implementation Strategies

PREMIS conformance degrees of freedom

Explicit Recording of Information: A repository is not required to explicitly record in its metadata management system the information populating a particular PREMIS semantic unit that it has implemented. However, this information must be recoverable in some way when it is needed (for example, to create an information package for exchange with another repository).

Use of Controlled Vocabularies: A repository is free to use (or not use) controlled vocabularies to populate PREMIS semantic units. If the repository chooses to use controlled vocabularies, it is free to use either internally defined vocabularies, or externally-defined, standardized vocabularies.

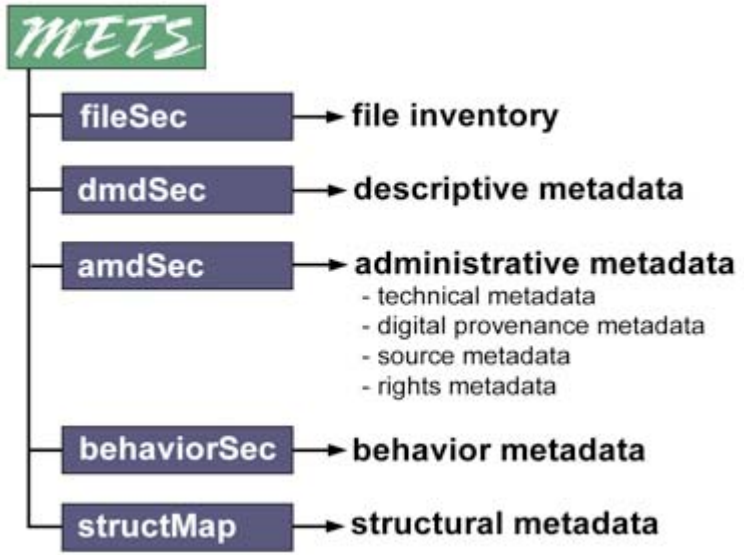
- Preservation Vocabularies

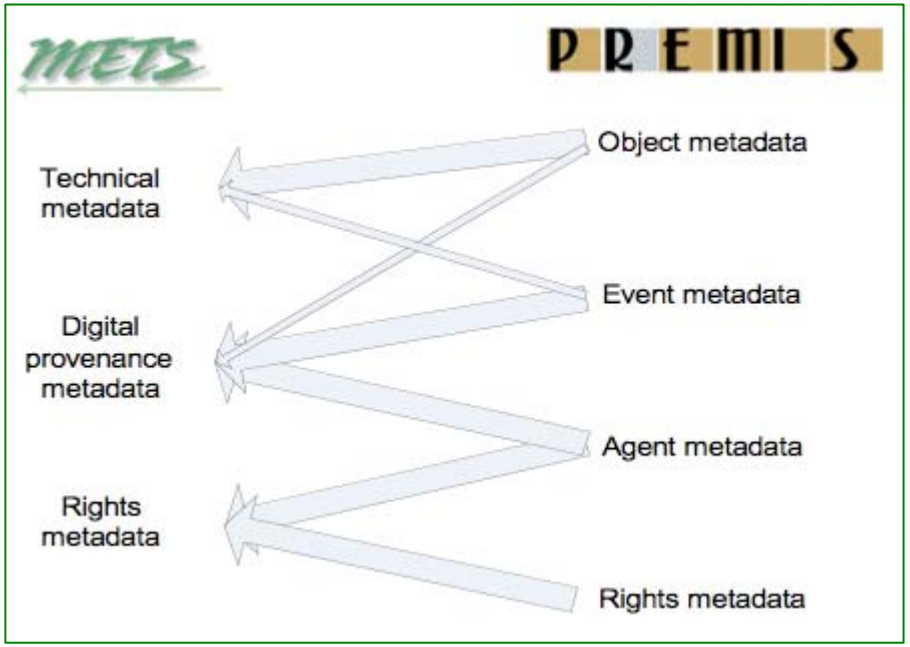
(see http://id.loc.gov/)

Preservation Events
Preservation Level Role
Cryptographic Hash Functions

PREMIS PREservation Metadata Implementation Strategies

METS/PREMIS relationships





Mapping PREMIS entities to METS metadata sections. Thick arrows show applicable subsection in METS for the named PREMIS entities; the thin arrow shows links from one PREMIS entity to another METS subsection.

http://www.dlib.org/dlib/july08/guenther/07guenther.html

PREMIS PREservation Metadata Implementation Strategies

Supporting documentation

Let´s take a tour on: http://www.loc.gov/standards/premis/

PREMIS WEB SITES, TOOLS AND E-MAIL

PREMIS maintenance activity Web site:
http://www.loc.gov/standards/premis/.

PREMIS Implementers' Group discussion list: pig@loc.gov.

To subscribe, send e-mail to
listserv@loc.gov with the message, "subscribe pig [your name]“

Please send comments and questions to premis@loc.gov.