wePreserve

Digital Preservation Metadata

Angela Dappert
The British Library, Planets, PREMIS EC
Barcelona
March 2009

Some of the slides on PREMIS are based on slides by
Priscilla Caplan, Florida Center for Library Automation
Rebecca Guenther, Library of Congress
Brian Lavoie, OCLC

wePreserve

Overview

- **Introduction to Digital Preservation Metadata**
  - What is Digital Preservation Metadata
  - Hands-on Exercise
  - Case Study: eJournals (1)
- **Preservation Metadata in Practice**
  - Workflow Issues
  - Tools and Standards
  - PREMIS Data Dictionary
    - Overview
    - Hands-on Exercise
    - Implementation Issues
  - Case Study: eJournals (2)

Overview

- **Introduction to Digital Preservation Metadata**
  - What is Digital Preservation Metadata
  - Hands-on Exercise
  - Case Study: eJournals (1)
- **Preservation Metadata in Practice**
  - Workflow Issues
  - Tools and Standards
  - PREMIS Data Dictionary
    - Overview
    - Hands-on Exercise
    - Implementation Issues
  - Case Study: eJournals (2)

Domain


Click and drag to move the glass
Help
Pages 17 and 18
Birdsong
Text
Audio
Magnify

What is Digital Preservation Metadata?

- **Metadata = data about data**
- **Information that is essential to ensure long-term accessibility of digital resources**

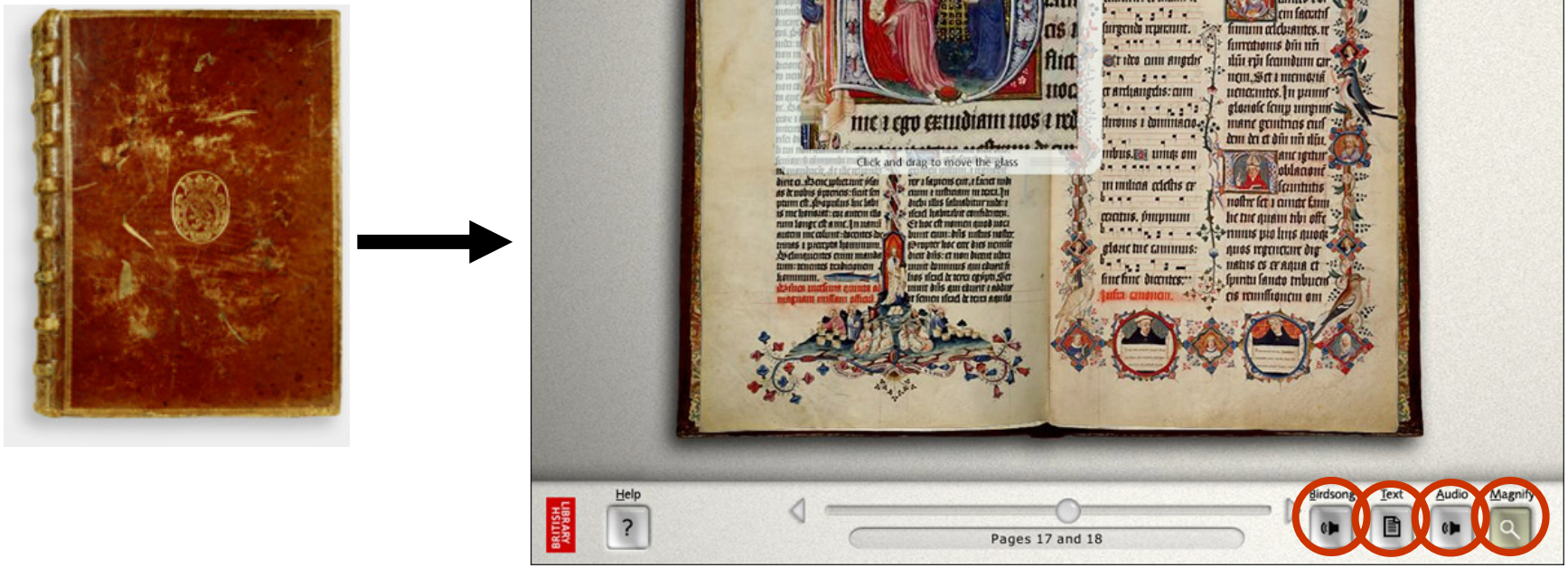
What is Digital Preservation Metadata?

- **A best guess on the future**
  - little experience with digital objects
  - uncertain future technical possibilities
  - uncertain future legal framework in which we will operate
- **Digital objects must be self-descriptive**
- **Must be able to exist independently from the systems which were used to create them**
  - XML (machine and human readable)

Why do we need new forms of metadata?

- Use Cases

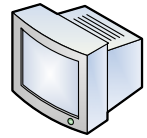
Supporting New Features

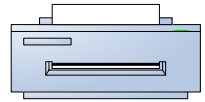


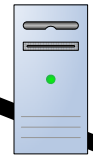
- **MetaD:**
  Semantic Information for the designated community

Technology Dependence

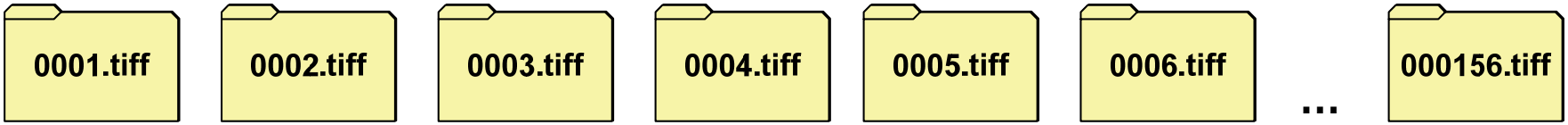
digital







Complex environments



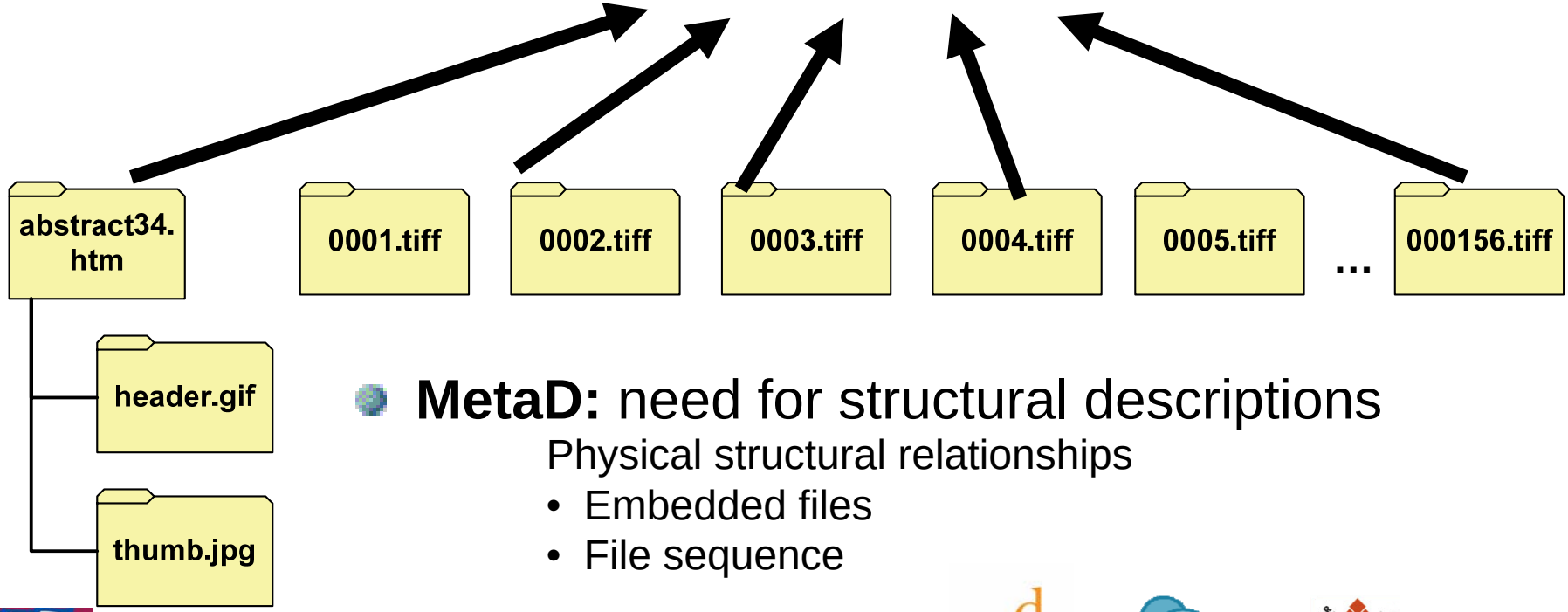
No direct access

- Not self-descriptive
- Complex formats

Technology Dependence

- **MetaD:**
  - need for detailed rendering information
    - Software
    - Hardware
    - Other dependencies: schemas, style sheets, encodings, etc.
  - need for format information

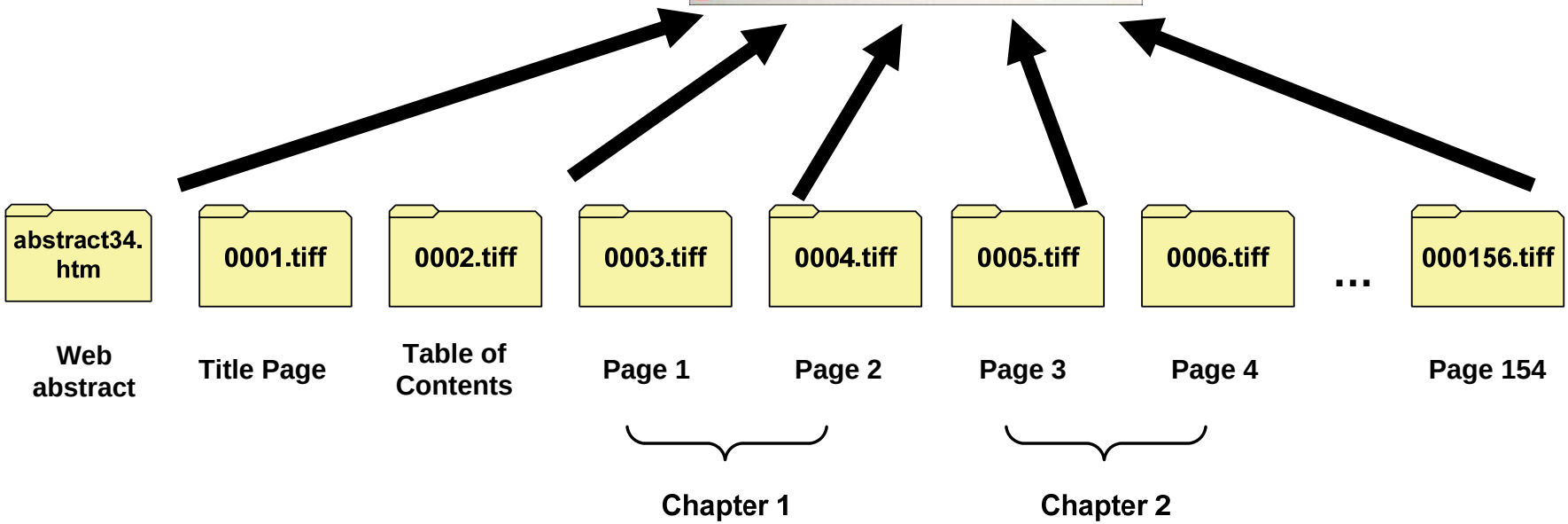
Complex Structures



- **MetaD:** need for structural descriptions
  Physical structural relationships
  - Embedded files
  - File sequence

Complex Structures

- **MetaD:** need for structural descriptions
  - Logical structural relationships



Mutability

- Intentional or accidental change
- Decay: rapid and potentially complete

Mutability

Viability: **the object is readable**

- **Action:**
  - Sound storage management practices, including climate control
  - Choice of resilient file formats
  - Media refreshment
    (copying data from one storage device to another)

- **MetaD:**
  - Data carrier metadata
    - type of medium
    - its preservation characteristics
    - age of medium
    - date of recording
    - usage patterns

Mutability

Fixity**: the object is unchanged**

- **Action:**
  - Regularly compute checksums (>=2)
- **MetaD:**
  - Checksums, message digests
  - Event creating them
    - Hash algorithms creating them
    - Date/Time
    - Originator

Mutability

Integrity**: the object is whole and unimpaired**

- **Action:**
  - format identification and validation
  - structural information:
    all files are there
    all files are named correctly
- **MetaD:**
  - event information for format identification and validation events
  - structural metadata

Mutability

Authenticity**: the object is what it purports to be**

- **Action:**
  - Procedural: virus protection
firewalls
tight authentication
intrusion detection
immediate attention to security alerts
  - Technical: replication
digital signatures
- **MetaD:**
  - Provenance metadata (events and agents)
  - Digital signatures
  - Access rights

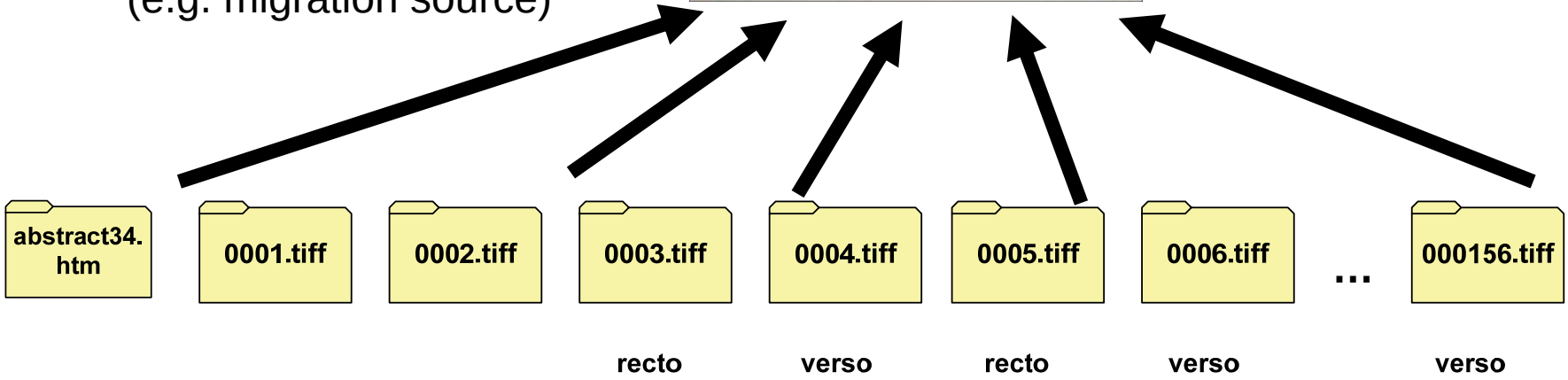
Obsolescence

- **Action:**
  - Frequent, pre-emptive preservation actions (migration, emulation)
  - During copyright period
  - Potential loss of object characteristics

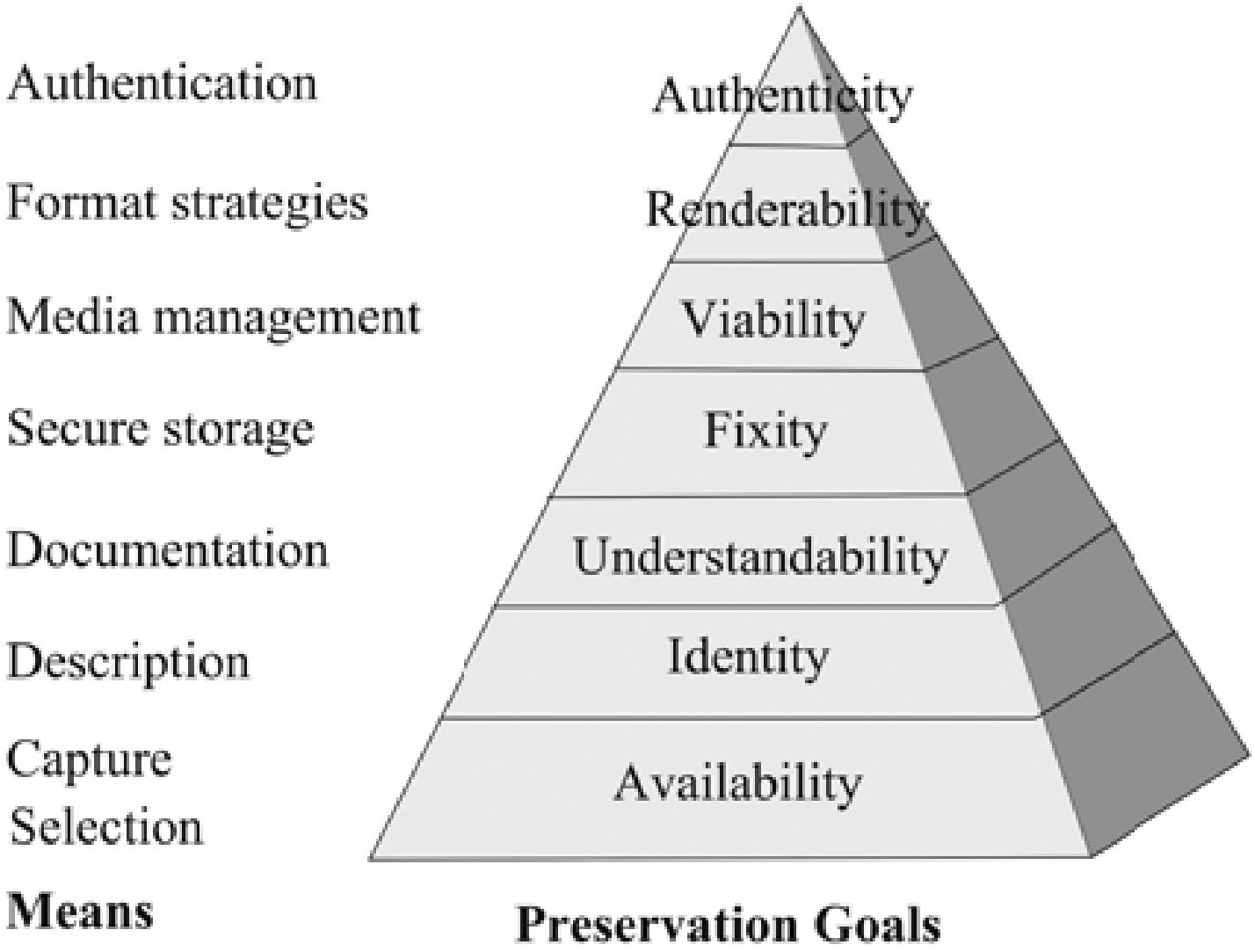
- **MetaD:**
  - Provenance metadata: history of all actions performed on the resource + custodianship
    - changes and decisions
    - dates
    - agents (decision maker + tools used)
  - Preservation action rights information
  - Significant properties

Context Descriptions

- **MetaD:**
  need for
  context descriptions
  - Original source
  - Related items
  (e.g. migration source)



Preservation Pyramid
(from Priscilla Caplan)



Conventional metadata categories with new functions

- **Administrative:**
  - NOW: vendor, shelf marks, submission information
  - technical download or upload information,
  - technical format specification of directory structures, files, schemata
  - …
- **Descriptive :**
  - greater granularity, descriptive information on more levels: journal -> issue -> article -> words (indexing for search engines) -> images contained, etc.;
  - more granularity in rights;
  - less subject information (?) because of automated search;
  - more facet information (?) to support faceted search

Conventional metadata categories with new functions

- **Structural:**
  - Now: issue belongs to journal, song is part of CD, …
  - greater granularity -> logical structural relationships; physical structural relationships,
  - …
- **Technical**
  - Now: dimensions of book, …
  - size in bytes,
  - number of files,
  - file formats,
  - environment,
  - hash functions
  - …

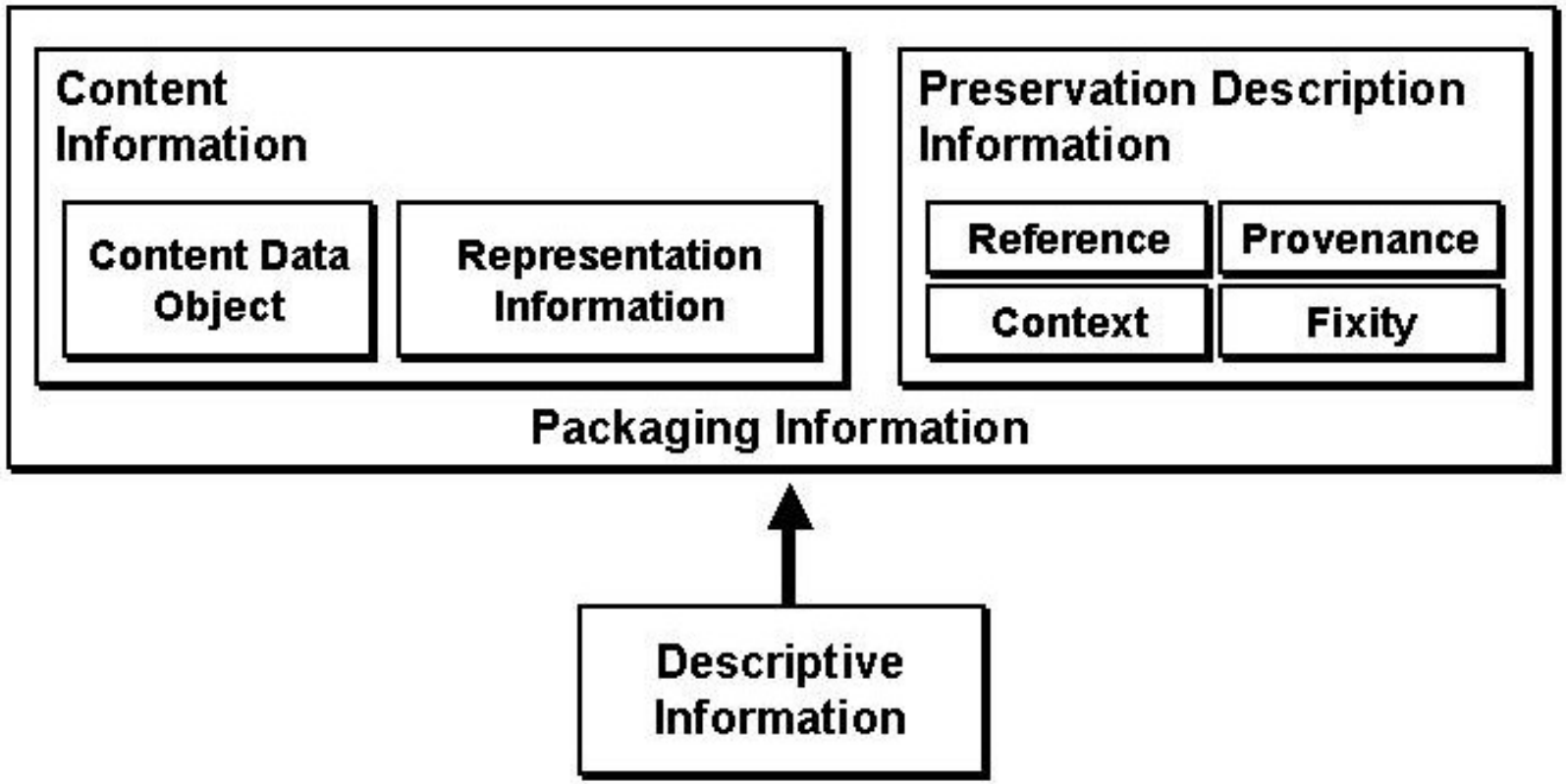
Conventional metadata categories with new functions

- **Metadata needed for resource discovery and delivery (e.g. structural information, rendering environment information) is also preservation metadata.**

- **Additionally**
  - made explicit for future generations
  - anticipating future uses
  - anticipating what will go wrong with the digital object

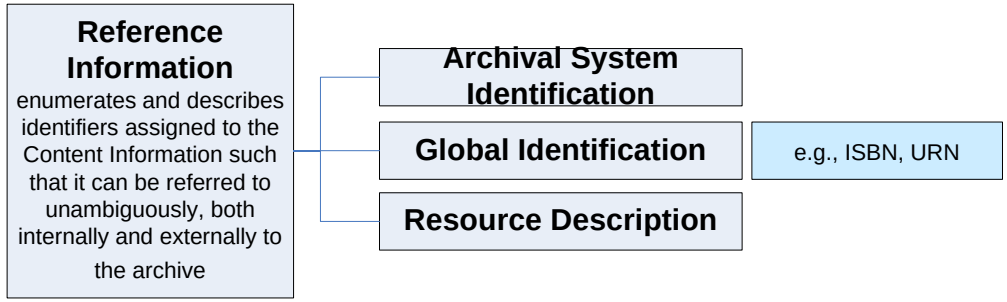
Preservation Metadata Categories - OAIS

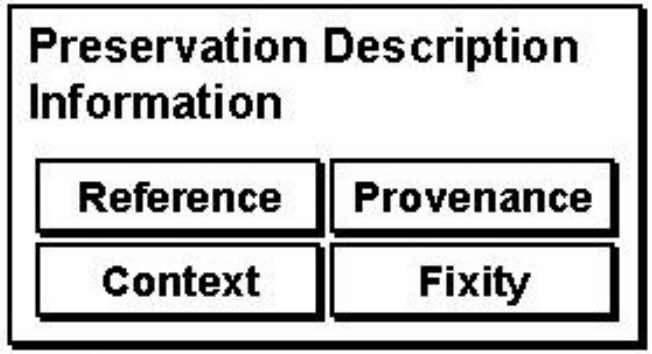
OAIS Model Preservation Metadata Categories



OAIS Model preservation metadata categories

- **Preservation Description Information**
  - Reference Information: to enumerate and describe identifiers
  - Provenance Information: to document the history of the content information (creation, modification, custody)
  - Context Information: to document the relationship of the content to its environment
  - Fixity Information: to document authentication mechanisms





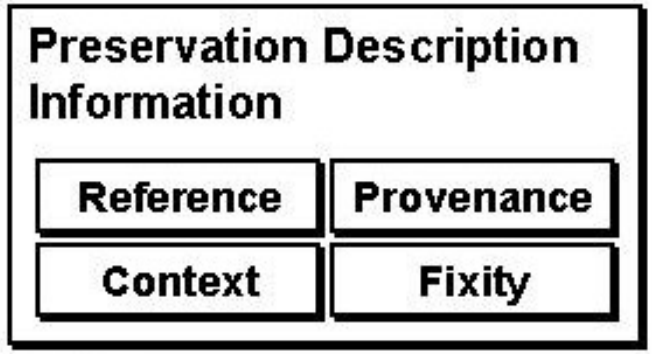
OAIS Model preservation metadata categories

- **Preservation Description Information**
  - Reference Information: to enumerate and describe identifiers
  - Provenance Information: to document the history of the content information (creation, modification, custody)
  - Context Information: to document the relationship of the content to its environment
  - Fixity Information: to document authentication mechanisms

Provenance Information
documents the history of the Content Information

Ingest Process History
- Institution
- Event Date/Time
- Event Type
- Event Description

Preservation History
- Institution
- Action Date/Time
- Action Type
- Action Description
- Technical Device

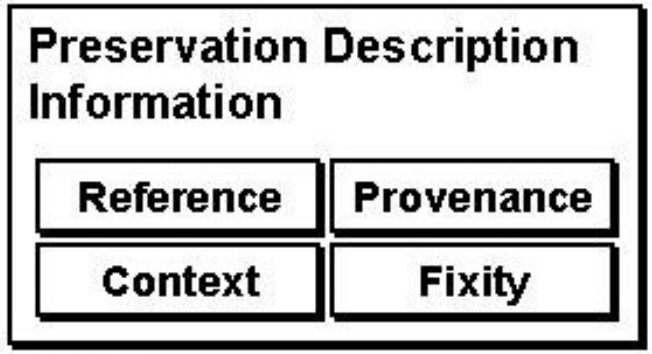


OAIS Model preservation metadata categories

- **Preservation Description Information**
  - Reference Information: to enumerate and describe identifiers
  - Provenance Information: to document the history of the content information (creation, modification, custody)
  - Context Information: to document the relationship of the content to its environment
  - Fixity Information: to document authentication mechanisms

Context Information
documents the relationships of the Content Information to its environment (e.g., why it was created, relationships to other Content Information)

- Reason for Creation
- Is Version Of
- Has Version
- Is Replaced By
- Replaces (migration)
- Is Required By
- Requires
- Is Part Of
- Has Part
- Is Referenced By
- References
- Is Format Of
- Has Format
- Same Intellectual Content As

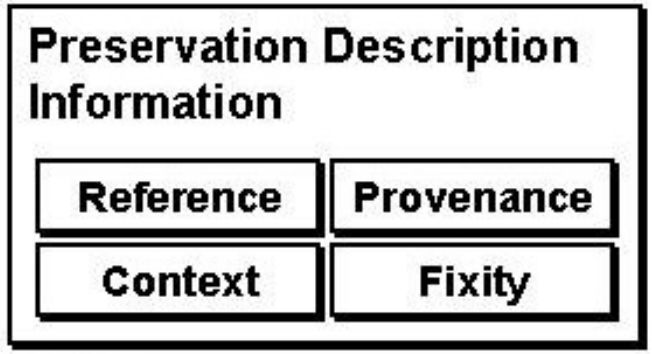


OAIS Model preservation metadata categories

- **Preservation Description Information**
  - Reference Information: to enumerate and describe identifiers
  - Provenance Information: to document the history of the content information (creation, modification, custody)
  - Context Information: to document the relationship of the content to its environment
  - Fixity Information: to document authentication mechanisms

Fixity Information
information validating the authenticity of the content information

Authentication
- Dig. Signature / Watermark / Time Stamp
- Checksum
- Encryption
- Documentation of Auth. Mechanism

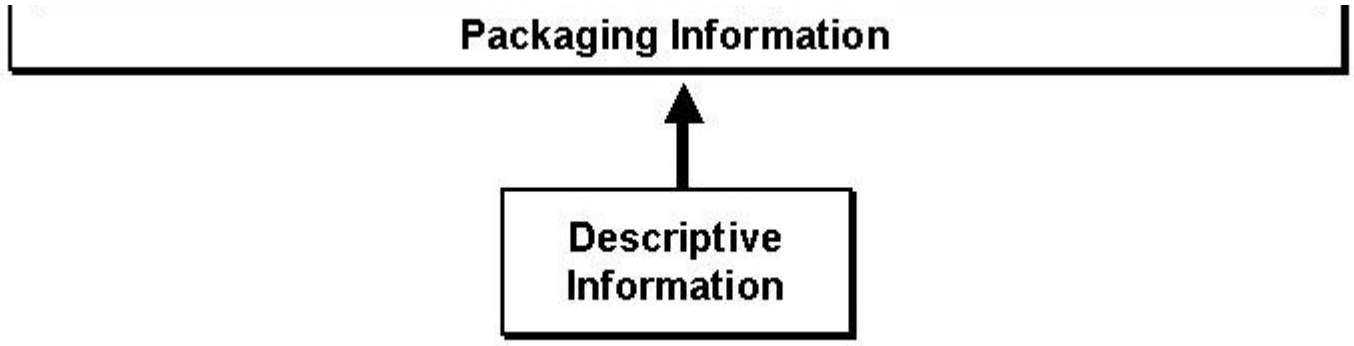


OAIS model categories

- **Packaging Information**
- **Descriptive Information the information used to aid searching, ordering, and retrieval of the objects**

Packaging Information

binds the digital object and its associated metadata into an identifiable unit or package (i.e., an Archival Information Package)

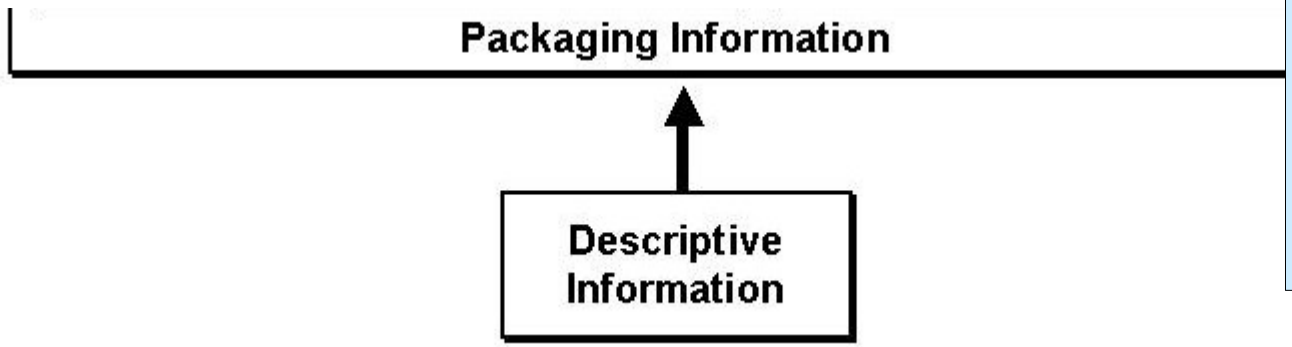


OAIS model categories

- **Packaging Information**
- **Descriptive Information the information used to aid searching, ordering, and retrieval of the objects**

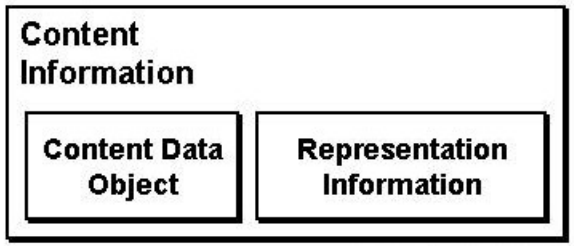
Descriptive Information
that helps users of the archive to locate and access information of potential interest.

DC.Title
DC.Creator
DC.Subject
DC.Description
DC.Publisher
DC.Contributor
DC.Date
DC.Type
DC.Format
DC.Identifier
DC.Source
DC.Language
DC.Coverage



OAIS model categories

- **Content Information**
  - Content Data Object
  - Representation Information the information needed for proper rendering, understanding, and interpretation of a content data object



Semantic Information
provides additional meaning for the interpretation of the content

E.g. three numbers interpreted as a date
E.g. the text is in English.

Structure Information
interprets the bits by organizing them into specific data types, groups of data types, and other higher-level meanings.

E.g. identify a bit stream as ASCII text characters,

Content Data Object Description
detailing the characteristics and features of the Content Data Object itself that are necessary to render and understand its content.

Directory structure and file naming conventions
Content type
Component types and their relationships
File description
Installation requirements
Size
Access inhibitors
Access facilitators
Significant properties
Functionality
Description of rendered content
Quirks
Documentation

Environment Description
describes a hardware/software environment capable of rendering or displaying the Content Data Object in the form in which it currently exists in the archival store.

- **Hardware Environment**
  - Computational Resources
  - Storage
  - Peripherals
- **Software Environment**
  - Rendering Programs
  - Operating System

Overview

- **Introduction to Digital Preservation Metadata**
  - What is Digital Preservation Metadata
  - Hands-on Exercise
  - Case Study: eJournals (1)
- **Preservation Metadata in Practice**
  - Workflow Issues
  - Tools and Standards
  - PREMIS Data Dictionary
    - Overview
    - Hands-on Exercise
    - Implementation Issues
  - Case Study: eJournals (2)

Hands-On Exercise

- **Imaginary eJournal submission (inspired by the Elsevier ScienceServer specification)**
- **You want to collect this content-type in your repository to ensure long-term access.**
- **It is the first time that you see this publisher's format and you start to think about your metadata needs.**

Hands-On Exercise

Goal:

- **Store metadata in the repository with the content to create complete, self-descriptive units**
- **Specify metadata profiles for archival information packages (AIP)**

Creating Metadata Profiles

1. **Which objects do we describe?**
   a. Which?
   b. How many?
2. **Which metadata do we need?**
   a. Which do we need?
   b. Which do we get?
3. **Which standard do we use for which metadata?**

Creating Metadata Profiles for eJournals

Answers are based on analysis of the

- **Concepts in the domain**
- **Sources of objects and metadata**
- **Technical properties of the repository**
- **Use Cases**
  - Functions supported (what is MetaD for?)
  - Workflow (how is MetaD used?)

Hands-On Exercise

- What sorts of digital objects need to be described?
- What are the relationships between them?
- What descriptive metadata can you find?
- Can you tell what events the objects have undergone?
- What technical metadata can you find?
- What information can you find that supports fixity, integrity and authenticity?
- What rights information can you find?

Don’t fret over details!

Overview

- **Introduction to Digital Preservation Metadata**
  - What is Digital Preservation Metadata
  - Hands-on Exercise
  - Case Study: eJournals (1)
- **Preservation Metadata in Practice**
  - Workflow Issues
  - Tools and Standards
  - PREMIS Data Dictionary
    - Overview
    - Hands-on Exercise
    - Implementation Issues
  - Case Study: eJournals (2)

Creating Metadata Profiles

1. **Which objects do we describe?**
   a. Which?
   b. How many?

- **What sorts of domain objects are you wanting to preserve?**
- **Do you want to describe intellectual entities, representations, files, bitstreams?**

Question 1a: Which objects do we describe?

- **For eJournals:**
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

Question 1a: Which objects do we describe?

- **For eJournals**:
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

January/February 2009
ISSN: 1082-9873
Vol. 15 No. 1/2
doi:10.1045/dlib.magazine

Question 1a: Which objects do we describe?

- **For eJournals:**
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

A Policy Checklist for Enabling Persistence of Identifiers
Nick Nicholas, Nigel Ward, and Kerry Blinco, *Link Affiliates*
doi:10.1045/january2009-nicholas

Question 1a: Which objects do we describe?

- **For eJournals**:
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

- provider specific
- XML
- HTML
- PDF

not identical content

Question 1a: Which objects do we describe?

- **For eJournals:**
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

Thumb.jpg in the XML representation

Question 1a: Which objects do we describe?

- **For eJournals**:
  - Journal
  - Issue
  - Article
  - Representation
  - File
  - Submission

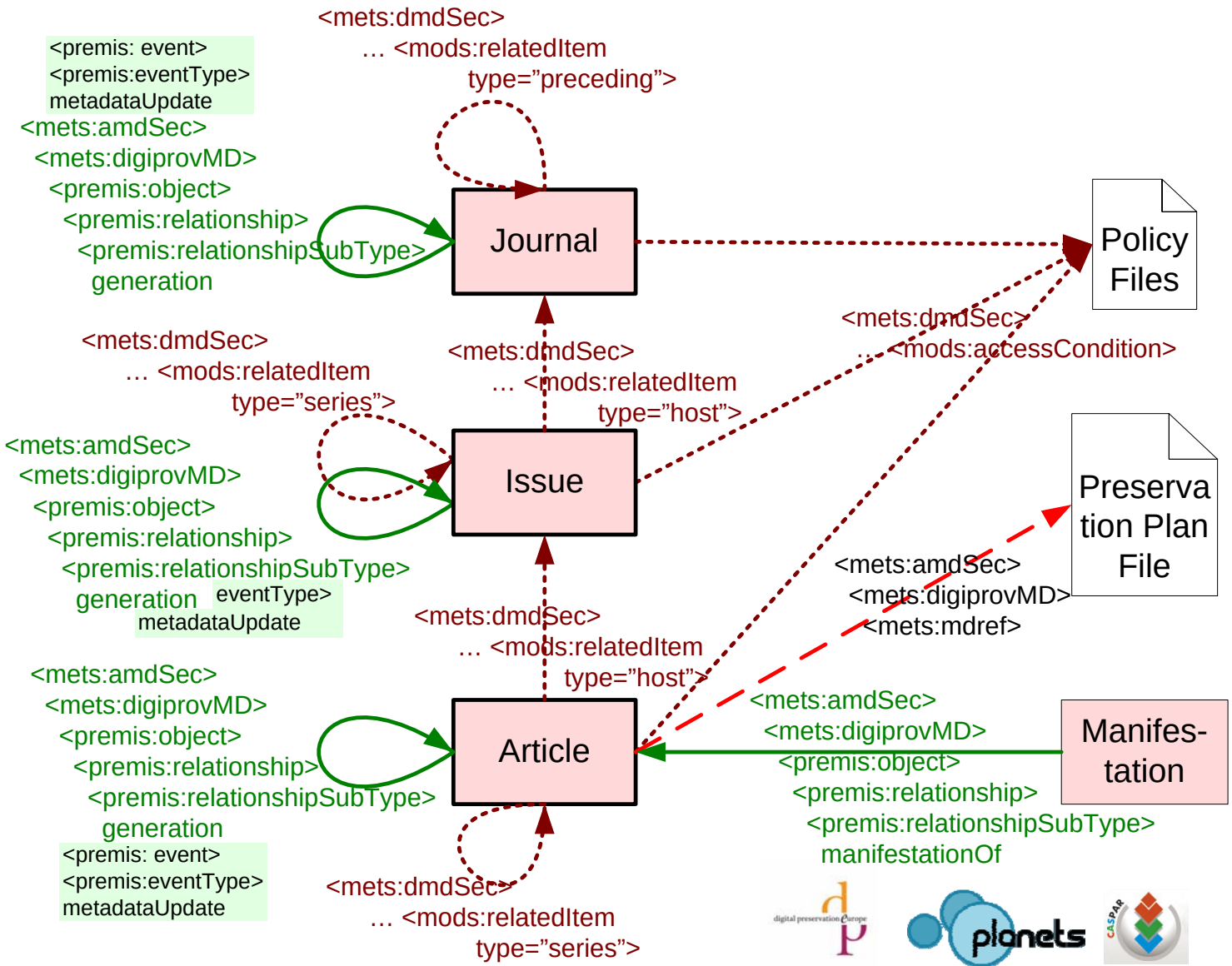
packages contain all the content files, metadata, manifests; for convenience, records provenance information (events) that are shared by many files

Question 1b : How many objects do we implement?

- **For eJournals:**

Because of the write-once architecture of the Digital Library System, we split objects into chunks which are updated together. This avoids, for example, creating new generations of journal objects with every submission of a new issue.

Example Diagram



Creating Metadata Profiles

2. **Which metadata do we need?**
    a. Which do we need?
    b. Which do we get?

Question 2a: Which metadata do we need?

- **Which functions are supported by the system and what information do they need?**
  - For how long do you want to retain the digital objects?
  - How intensive are your preservation needs?
  - What technical metadata do you need to record to perform your business processes?
  - What other metadata do you need to record to perform your business processes?
  - How diverse is your user base? Does this influence your preservation needs?
  - How self-documenting are your digital objects?

Question 2a: Which metadata do we need?

- **Which functions are supported by the system and what information do they need?**
  - Can the repository demonstrate the fixity, integrity, authenticity of archived materials?
  - What preservation strategies (migration, normalization, emulation, cannonicalization, etc.) will the system implement; how will it use metadata in this process?
- **Which relationships exist between objects?**
- **Which events, agents, rights do we describe?**
  - Which of these events change the objects or their metadata?

Question 2a: Which metadata do we need?

For eJournals:

- **Which functions are supported by the system and what information do they need?**

  Preservation, technical requirements, resource discovery, management information, reading room, …

  Preservation metadata does not exist in isolation!

Question 2a: Which metadata do we need?

For eJournals:

- **Which relationships exist between objects?**
  - generation, part-of, host, migrated-from, series, preceding, manifestation-of, …
- **Which events, agents, rights do we describe?**
  - Accession, validation, virus check, uncompress, metadata extraction, format identification, migration, …

Question 2b: Which metadata can we get?

For eJournals:

- Many suppliers of eJournals to one repository
- Formats of metadata and content are out of the control of the repository
- Translators to the internal metadata format need to be written
- To guide the writing of translators, the metadata profiles need to be very precise so that the translators will produce high-quality, uniform metadata

Overview

- **Introduction to Digital Preservation Metadata**
  - What is Digital Preservation Metadata
  - Hands-on Exercise
  - Case Study: eJournals (1)
- **Preservation Metadata in Practice**
  - Workflow Issues
  - Tools and Standards
  - PREMIS Data Dictionary
    - Overview
    - Hands-on Exercise
    - Implementation Issues
  - Case Study: eJournals (2)

Creation of Metadata

- A common metadata framework, used by both the producer and the repository is advantageous.
- Repositories may have to normalize metadata.
- The actor who is closest to the information to be used as metadata creates it.

Creation of Metadata

- **Producer**
  - Events that occur before ingest into the repository
  - Technical information about the creation of the object
  - Fixity Information
  - Context Information
  - Representation Information
  - Significant Properties
  - Intellectual Property rights

Creation of Metadata

- **Repository**
  - Extracted technical information (JHOVE, NLNZ extraction tool)
  - Extracted structural information (METAe)
  - Registries
  - Events at ingest, migration and other points in the life-cycle
  - Significant Properties

From Creation to Repository

- **Negotiation: Submission agreement between producer and repository**
  - Means of transmission
  - Verification process
  - Formats and standards
  - Process by which the repository can request re-transmission
- **Files should be verified**
  - against checksums sent by the producer
  - with the help of characterisation tools

Manual or Automatic Creation of Metadata

- **Some metadata needs to be created by hand**
- **Automatic production of metadata is the goal**
  - Higher granularity of description increases the number of objects to be described
  - populated by ingest software
  - extracted by tools
    - JHOVE, NLNZ Metadata Extraction Tool

Metadata Registries

- **Help interoperability , exchange**
- **Help metadata reuse**
- **Examples**
  - file format: Pronom, GDFR
  - environment information: Pronom

    e.g. for any file format, give list of software that can create, render, edit, identify, validate, extract metadata
  - object properties and their extraction: Pronom, XCEL
  - controlled vocabulary for values: Library of Congress

Metadata Storage

- **Discrete files**
  - simplest
  - text files (often using XML tagging)
  - associated with the digital objects by persistent IDs
- **Database management system (relational, object-oriented, or XML)**
  - fast access, easy update, and ease of use for query and reporting
  - capable of storing a relational model of complex objects
  - requires a higher level of technical commitment
- **Embed metadata in the objects themselves**
  - possible with some file formats

Metadata Storage

- **Metadata stored with the content data files?**
  - Makes digital objects self-descriptive
  - Harder to separate the metadata from the content
  - The same preservation strategies that are applied to the content can be applied to the metadata.

Metadata Storage

- **Do you need to store metadata or can it be extracted on the fly?**
  - For what is it used? Do you need to search by it?
  - Will you know how to extract it later?

Updating Metadata

- **Record update as Provenance Information**

Overview

- **Introduction to Digital Preservation Metadata**
  - What is Digital Preservation Metadata
  - Hands-on Exercise
  - Case Study: eJournals (1)
- **Preservation Metadata in Practice**
  - Workflow Issues
  - Tools and Standards
  - PREMIS Data Dictionary
    - Overview
    - Hands-on Exercise
    - Implementation Issues
  - Case Study: eJournals (2)

Some Metadata Tools

- **Characterisation**
  - file format identification, validation, characterisation and assessment
  - JHOVE, DROID, XCEL characteriser
- **Registries**
  - file formats, environments, properties, property extraction for file formats, controlled vocabularies
  - PRONOM, GDFR, XCEL
- **Metadata creation and transformation**
  - see http://www.loc.gov/standards/premis/tools.html
  - http://www.loc.gov/standards/mets/mets-tools.html

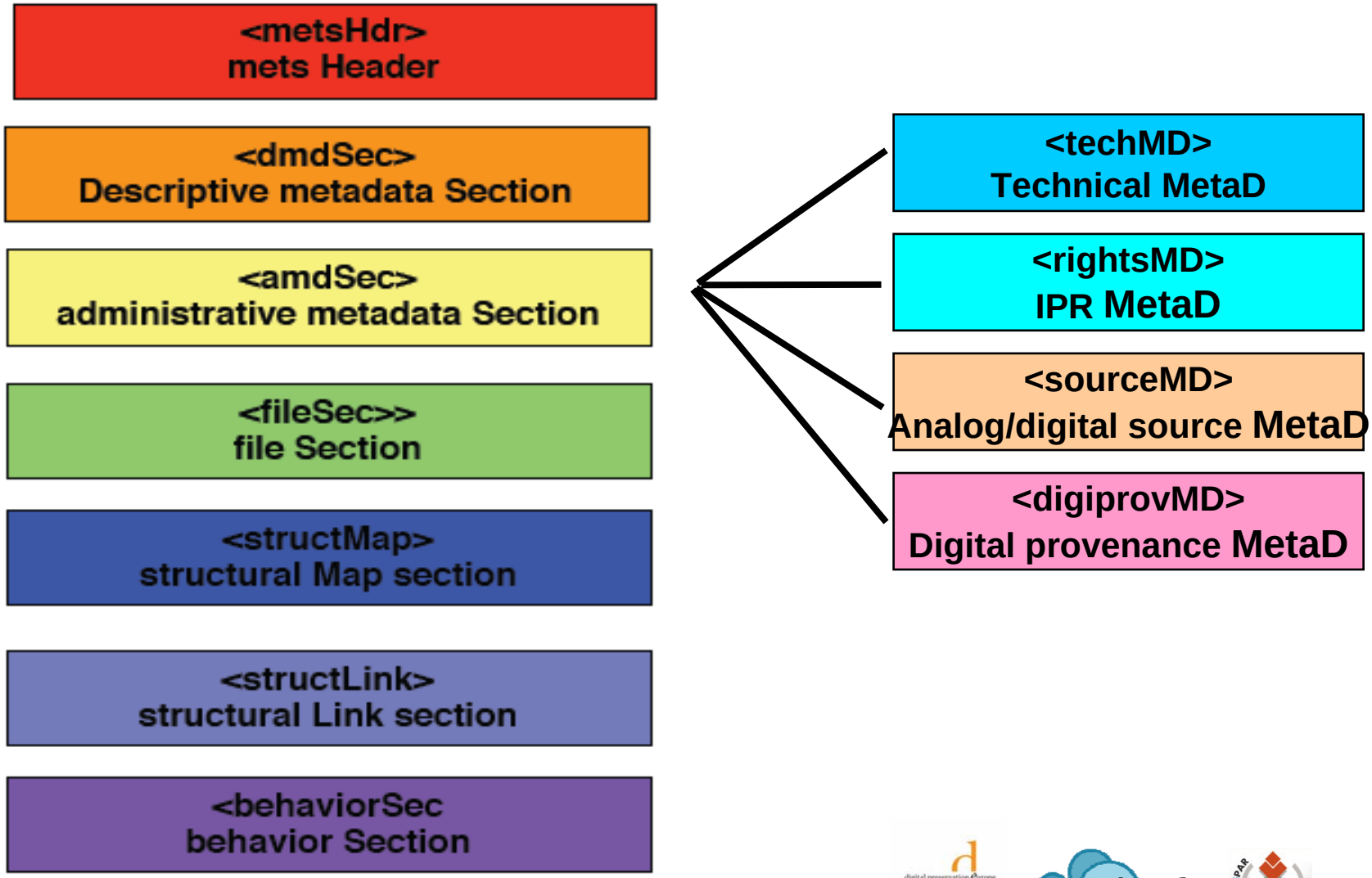
Some Metadata Standards

- **Descriptive metadata**
  - Dublin Core
  - Metadata Object Description Schema (MODS)
  - MARCXML MARC 21 Schema (MARCXML)
  - VRA Core (description of works of visual culture as well as the images that document them)
- **Content-type specific technical metadata**
  - textMD Schema for Technical Metadata for Text
  - MIX NISO Technical Metadata for Digital Still Images
- **Core preservation metadata**
  - PREMIS

Metadata Containers

- Often XML based
- Encapsulates administrative, structural, and descriptive metadata about digital objects
- Extensible:
  elements from other schemas can be plugged in
- Records the structure of digital objects, and the names and locations of the files that comprise those objects.
- Records relationships among the metadata and among the pieces of the complex objects

METS Container



Inserting Technical Metadata in a METS Document

```
<mets>
 <amdSec>
  <techMD>
   <mdWrap>
    <xmlData>
        <!--  insert data from different namespace here  -->
    </xmlData>
   </mdWrap>
  </techMD>
 </amdSec>
 <fileSec />
 <structMap />
</mets>
```

Metadata Containers (cntd.)

- Describes and attaches executable behaviour appropriate for content
- A unit of storage (OAIS AIP)
  or a transmission format (OAIS SIP or DIP)
- Content-type independent

- Batch processing for creation, processing, retrieval, and presentation
- Text editor, XML editor, or a forms-based user interface built and customized to your collections and to your working environment

Metadata Containers

- METS: Metadata Encoding and Transmission Standard
- MPEG-21: Digital Item Declaration Language (DIDL)
- Fedora Object XML (FOXML)
- XFDU
- IMS Content Packaging Specification (IMS-CPS)
- Sharable Content Object Reference Model (SCORM)
- CCSDS XML Packaging Approach in the ESA Data Disposition System
- WARC File Format
- Open Archives Initiative Object Reuse and Exchange
- RAMLET

Preservation Metadata Element Sets

- RLG/OCLC Working Group's A Metadata Framework to Support the Preservation of Digital Objects (=> PREMIS)
- OCLC's Digital Archive Metadata Elements
- The National Library of Australia
- The National Library of New Zealand's Metadata Standards Framework
- Cornell University Library Proposed Metadata Elements

- LMER Long-term Preservation Metadata for Electronic Resources
- PREMIS *Preservation Metadata: Implementation Strategies*

Preservation Metadata Element Sets

- **Implementable preservation metadata**
  - rigorously defined
  - supported by guidelines/recommendations for management and use
  - emphasis on automated workflows
- **Core preservation metadata**
  - Relevant to a wide range of digital preservation systems and contexts
  - What most preservation repositories need to know to preserve digital materials over the long-term

Preservation Metadata Element Sets

- **“Technical neutrality”:**
  - Digital archiving system: no assumptions about specific archiving technology, system/DB architectures, preservation strategy
  - Metadata management: no assumptions about whether metadata is stored locally or in external registry; recorded explicitly or known implicitly; instantiated in one metadata element or multiple elements
  - Promotes flexibility, applicability in wide range of contexts

wePreserve

Overview

- **Introduction to Digital Preservation Metadata**
  - What is Digital Preservation Metadata
  - Hands-on Exercise
  - Case Study: eJournals (1)
- **Preservation Metadata in Practice**
  - Workflow Issues
  - Tools and Standards
  - PREMIS Data Dictionary
    - Overview
    - Hands-on Exercise
    - Implementation Issues
  - Case Study: eJournals (2)

Preservation Metadata Element Sets

The PREMIS Data Dictionary: Information you need to know for preserving digital documents

Preservation Metadata: Implementation Strategies

Scope

- **What PREMIS DD** is**:**
  - Common data model for organizing/thinking about preservation metadata
  - Guidance for local implementations
  - Standard for exchanging information packages between repositories

Scope

- **What PREMIS DD** is not**:**
  - Out-of-the-box solution: Choice of actual elements is driven by your business needs and documented in application profiles.
  - All needed metadata: excludes business rules, format-specific technical metadata, descriptive metadata for access, non-core preservation metadata, detailed agent metadata, intellectual entity metadata, information about the metadata itself (e.g., who obtained or recorded a value, when last changed...)
  - Lifecycle management of objects outside repository
  - Rights management: limited to permissions regarding actions taken within repository

Activities

- **Data Dictionary (PREMIS 2.0)**
  - http://www.loc.gov/standards/premis/v2/premis-2-0.pdf
- **Guidelines for using PREMIS with METS (draft available at:)**
  - http://www.loc.gov/standards/premis/premis-mets.html
- **PREMIS Implementers' Registry**
  - http://www.loc.gov/standards/premis/premis-registry.php

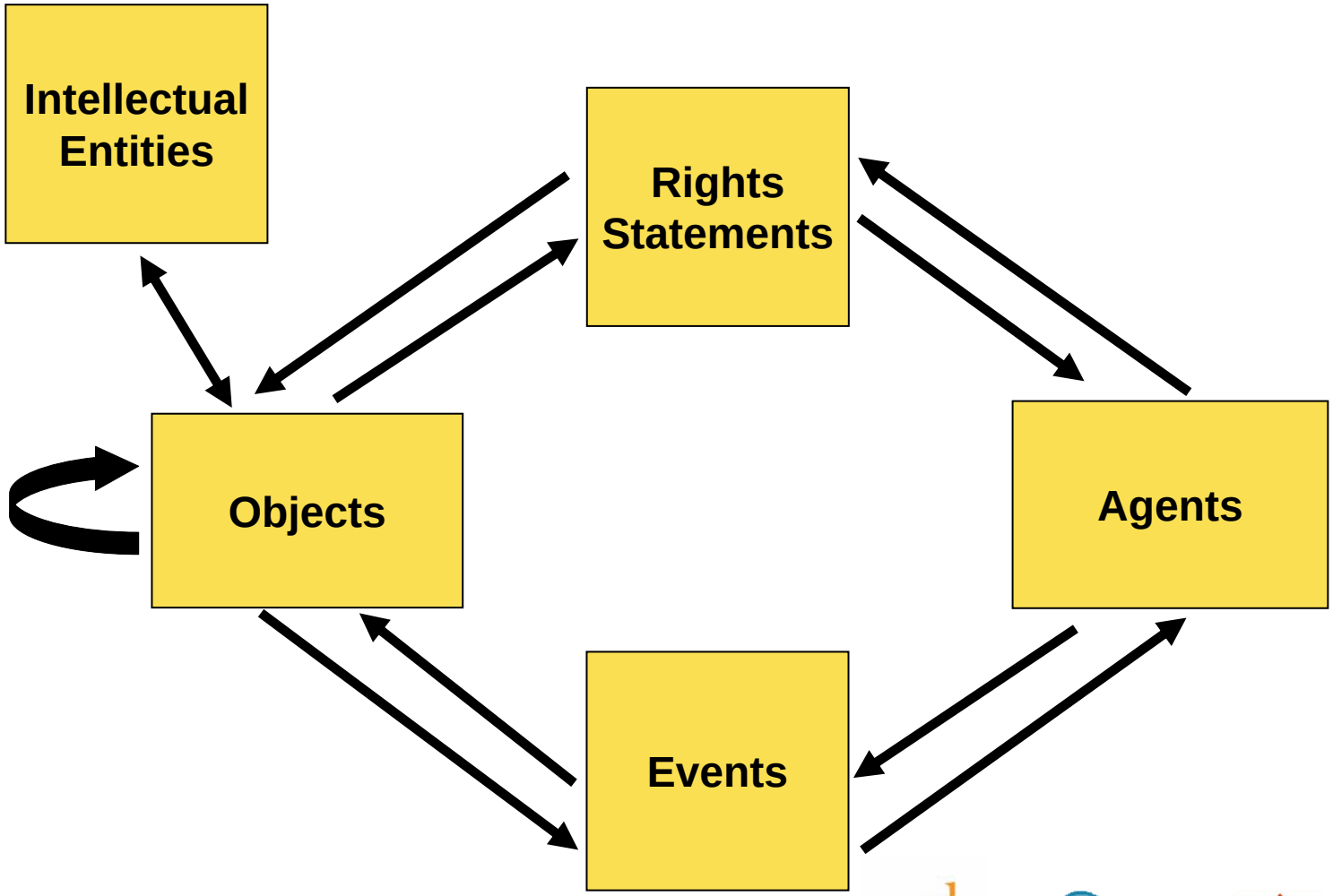
The PREMIS Data Model

- **Data model includes:**
  - Entities: “things” relevant to digital preservation that are described by preservation metadata (Intellectual Entities, Objects, Events, Rights, Agents)
  - Relationships between Entities
  - Properties of Entities (semantic units)

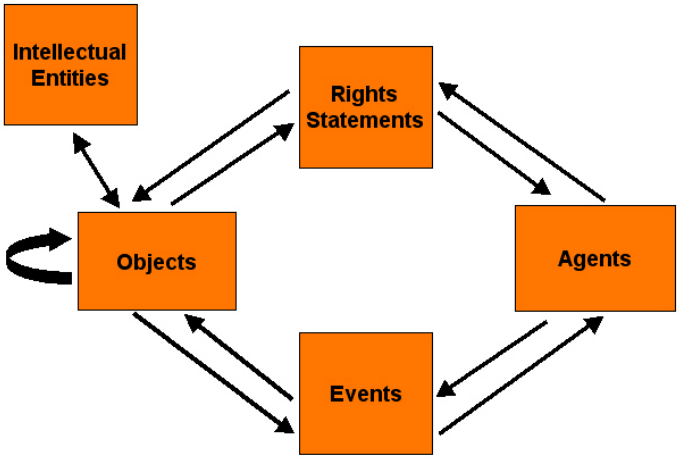
The PREMIS Data Model

- **Data model includes:**
  - Entities: “things” relevant to digital preservation that are described by preservation metadata (Intellectual Entities, Objects, Events, Rights, Agents)
  - Relationships between Entities
  - Properties of Entities (semantic units)

PREMIS Data Model



Intellectual Entities

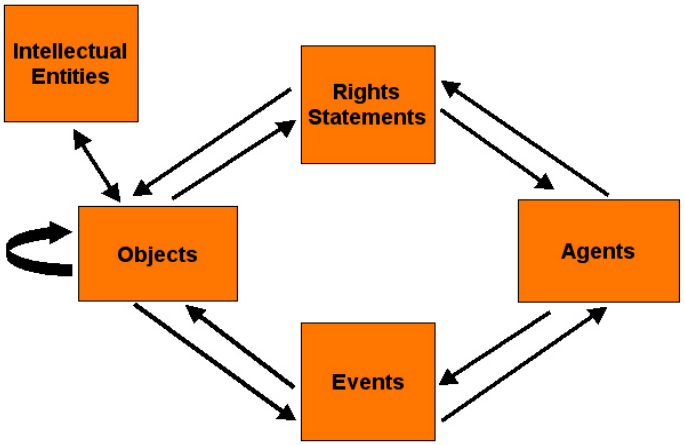


- Set of content that is considered a single intellectual unit for purposes of management and description (e.g., a book, a photograph, a map, a database)
- May include other Intellectual Entities (e.g. a website that includes a web page)
- **Has one or more digital representations**
- Not fully described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

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)

Objects

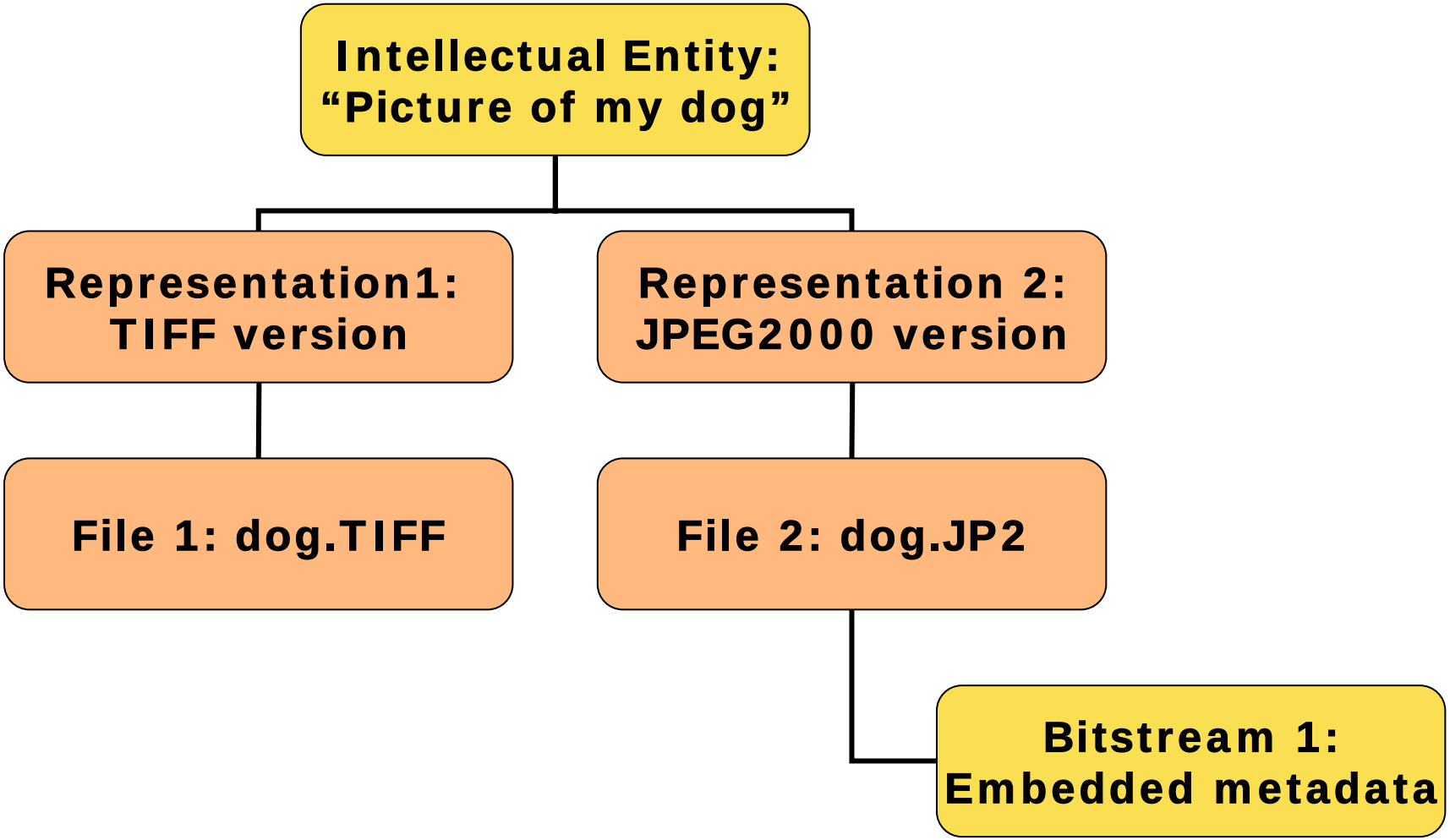


- **Discrete unit of information in digital form**
- **“Objects are what the repository actually preserves”**
- **Three types of Object:**
  - **FILE:** named and ordered sequence of bytes that is known by an operating system
  - **REPRESENTATION:** set of files, including structural metadata, that, 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)

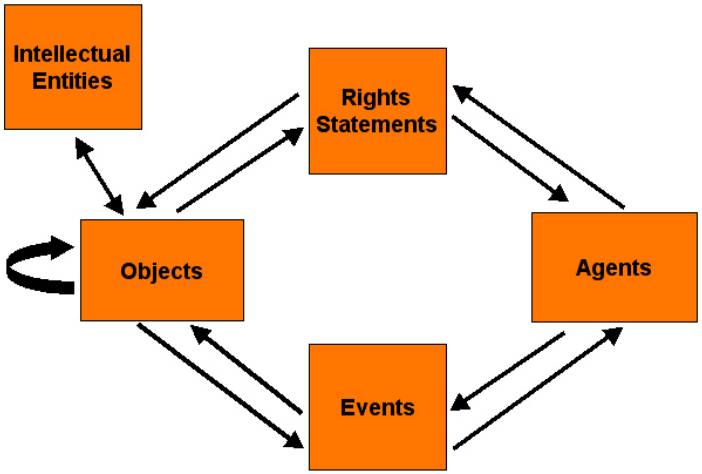
Examples:

- **chapter1.pdf (a file)**
- **chapter1.pdf + chapter2.pdf + chapter3.pdf (representation of a book w/3 chapters)**
- **TIFF file containing header and 2 images (2 bitstreams (images), each with own set of properties (semantic units): e.g., identifiers, technical metadata, inhibitors, …)**

Object Example: Photo in Two Formats



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

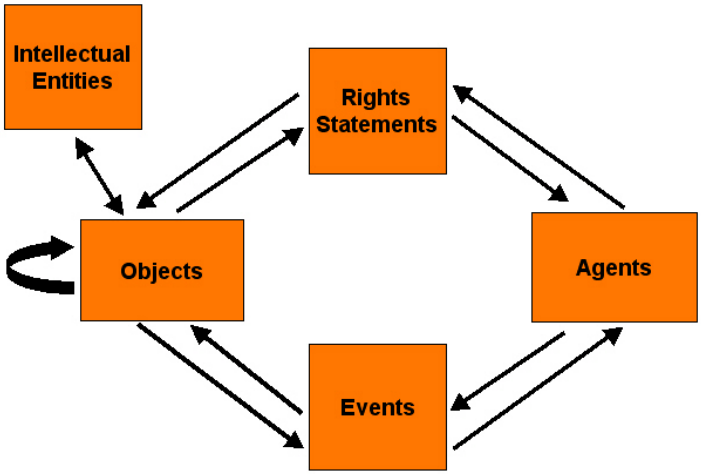
Examples:

- Validation Event: use JHOVE tool to verify that chapter1.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-to-date format

Some eventTypes

Capture	Compression
Deaccession	Decompression
Decryption	Deletion
Dig. signature validation	Dissemination
Fixity check	Ingestion
Message digest calculation	Migration
Normalization	Replication
Validation	Virus check

Agent

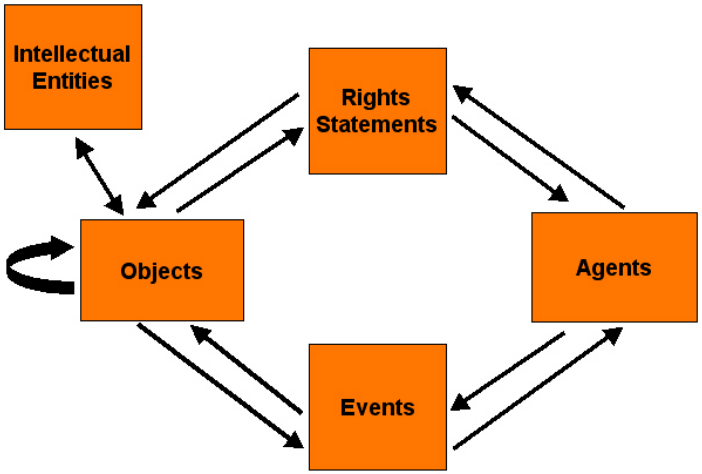


Examples:

- Priscilla Caplan (a person)
- Florida Center for Library Automation (an organization)
- Dark Archive in the Sunshine State implementation (a system)
- JHOVE version 1.0 (a software program)

- Person, organization, or software program/system associated with an Event or a Right
- Agents are associated only indirectly to Objects through Events or Rights
- Not defined in detail in PREMIS DD; not considered core preservation metadata beyond identification

Rights



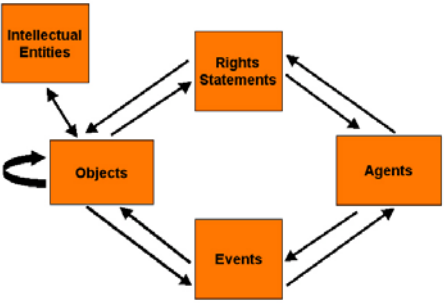
- **An agreement with a rights holder that grants permission for the repository to undertake an action(s) associated with an Object(s) in the repository.**
- **Not a full rights expression language; focuses exclusively on permissions that take the form:**
  - Agent X grants Permission Y to the repository in regard to Object Z.

Example:

- **Priscilla Caplan grants FCLA digital repository permission to make three copies of metadata_fundamentals.pdf for preservation purposes.**

The PREMIS Data Model

- **Data model includes:**
  - Entities: “things” relevant to digital preservation that are described by preservation metadata (Intellectual Entities, Objects, Events, Rights, Agents)
  - Relationships between Entities
  - Properties of Entities (semantic units)



serve

Relationships

- **PREMIS Data Dictionary supports expression of relationships between (see arrows):**
  - Different Objects
    - Across same level or different levels
  - Different Entities
- **Types of relationships:**
    - Structural: relationships between parts of a whole "A is part of B",
    - Derivation: relationships resulting from replication or transformation of an Object
      "A is scanned from B", "A is a version of B"
- **Relationships are established through reference to Identifiers of other Entities**

Relationships between Objects: Which, How, Why

- WHICH Objects are related?
- HOW are the Objects related?
- WHY are the Objects related?
  - Event?

Example: Structural relationship File “is part of” Representation

relationship [part of the description of File]

relationshipType = structural

relationshipSubType = is part of

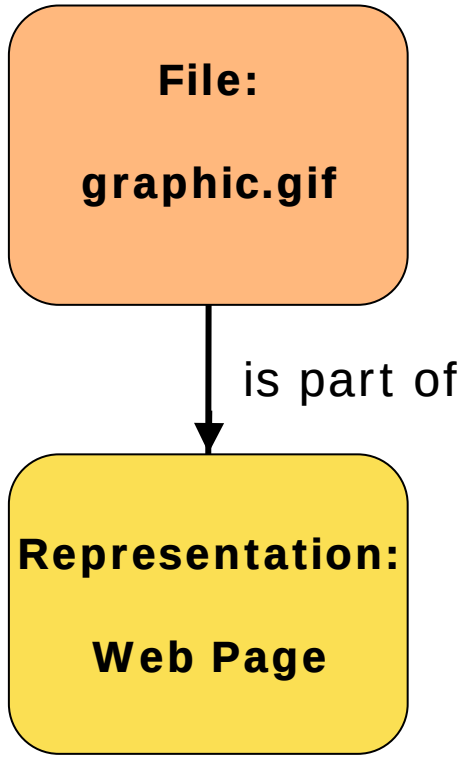
relatedObjectIdentification [the Web page]

relatedObjectIdentifierType = repositoryID

relatedObjectIdentifierValue = 0385503954

relatedObjectSequence = 0

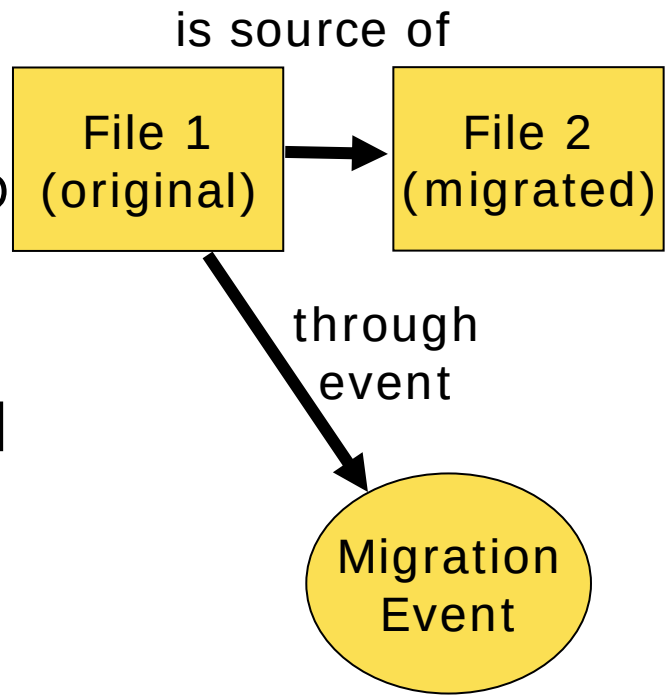
relatedEventIdentification [none]



Example: Derivation relationship File 1 “is source of” File 2

relationship [part of description of File 1]

- relationshipType = derivation
- relationshipSubType = is source of
- relatedObjectIdentification [identifier of File 2]
  - relatedObjectIdentifierType = repositoryID
  - relatedObjectIdentifierValue = F004400
  - relatedObjectSequence [none]
- relatedEventIdentification [Migration Event ID]
  - relatedEventIdentifierType = repEventID
  - relatedEventIdentifierValue = E0192
  - relatedEventSequence [none]



The PREMIS Data Model

- **Data model includes:**
  - Entities: “things” relevant to digital preservation that are described by preservation metadata (Intellectual Entities, Objects, Events, Rights, Agents)
  - Relationships between Entities
  - Properties of Entities (semantic units)

Semantic units

- **A semantic unit is a property of an Entity**
  - Something you *need to know* about an Object, Event, Agent, Right
- **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]**

Sample Data Dictionary Entry

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.		

Example: Object Entity

- **Main types of information**
  - identifier
  - technical object characteristics
  - creation information
  - software and hardware environment
  - digital signatures
  - relationships to other objects
  - links to other types of entity

Example: Object Entity

- **Main types of information**
  - identifier
  - technical object characteristics
  - creation information
  - software and hardware environment
  - digital signatures
  - relationships to other objects
  - links to other types of entity

Example: objectCharacteristics

- **Technical properties common to all/most file formats, not format specific**
- **Container for subunits:**
  - compositionLevel
  - fixity
  - size
  - format
  - creatingApplication
  - inhibitors
  - objectCharacteristicsExtension

Example: Object Entity

- **Main types of information**
  - identifier
  - technical object characteristics
  - creation information
  - software and hardware environment
  - digital signatures
  - relationships to other objects
  - links to other types of entity

Example Semantic Unit: Environment

- **What is needed to render or use an object**
  - Operating system
  - Application software
  - Computing resources

Environment example: ETD (PDF file)

- environmentCharacteristic:=known to work
- environmentPurpose=render
- software/swName=Adobe Acrobat Reader
- software/swVersion=6.1
- software/swType=renderer
- software/swDependency=Windows NT
- software/swName= Windows NT
- software/swVersion=5.0
- software/swType=operatingSystem
- hardware/hwName=Intel Pentium II
- hardware/hwType=processor
- dependency/dependencyName=Mathematica 5.2 True Type math fonts

Overview

- **Introduction to Digital Preservation Metadata**
  - What is Digital Preservation Metadata
  - Hands-on Exercise
  - Case Study: eJournals (1)
- **Preservation Metadata in Practice**
  - Workflow Issues
  - Tools and Standards
  - PREMIS Data Dictionary
    - Overview
    - Hands-on Exercise
    - Implementation Issues
  - Case Study: eJournals (2)

Object Exercise

1. Which PREMIS version was used for defining the object?
2. What is the identifier of the object?
3. Which significant property of the object must be preserved in a preservation action?
4. Which message digest algorithm was used to compute the checksum of the object?
5. What file format has the object
6. What is the corresponding registry code that was recorded and which registry was used to record it?
7. What software was used to create the object?
8. Which Extension Schema was used to record technical metadata?
9. On what data carrier is the object stored?
10. What software tools are recommended for rendering the object?
11. How many related items are recorded?
12. What is the nature of the relationships?
13. How many linking events have been recorded?
14. What do we know about them?

Event and Agent Exercise

1. What are the types of the 3 events?
2. What is the type of the agent?
3. Are there other agents captured in this information?
4. To what objects do the events link?
5. Are there other objects the events might link to?

Overview

- **Introduction to Digital Preservation Metadata**
  - What is Digital Preservation Metadata
  - Hands-on Exercise
  - Case Study: eJournals (1)
- **Preservation Metadata in Practice**
  - Workflow Issues
  - Tools and Standards
  - PREMIS Data Dictionary
    - Overview
    - Hands-on Exercise
    - Implementation Issues
  - Case Study: eJournals (2)

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

PREMIS and METS sections

- **You can't put all PREMIS metadata directly under amdSec**
- **What sections to use for PREMIS metadata?**
  - **Alternative 1**
    - Object in techMD
    - Event in digiProvMD
    - Rights in rightsMD
    - Agent with event or rights
  - **Alternative 2**
    - Everything in digiProvMD
  - **Alternative 3**
    - Everything in techMD
- **How many administrative MD sections to use?**

Elements Defined in Both METS and PREMIS

METS	PREMIS
METS: SIZE	PREMIS: size
METS: CHECKSUM, CHECKSUMTYPE	PREMIS: fixity
METS: MIMETYPE	PREMIS: <format>
METS ID/Idref: used to associate metadata in different sections and for different files	PREMIS identifiers: explicit linking between entity types
METS structMap: structural relationships, hierarchical, links the elements of the structure to content files and metadata	PREMIS <relationship>: all kinds of relationships, including structural

Should semantic units be recorded redundantly?

- **Options when there is overlap between PREMIS and another technical metadata schemas**
  - Record only outside PREMIS (e.g. in METS)
  - Record only in PREMIS
  - Record in both
- **Are there advantages in using PREMIS semantic units?**
- **Is it important to keep PREMIS metadata together as a unit? There may be an advantage for reuse and maintenance purposes**
- **Will there be problems synchronizing updates?**
- **Are they repeatable (e.g. attribute vs. element)?**
- **Are they granular (e.g. Software name and version separately or together)**

Overview

- **Introduction to Digital Preservation Metadata**
  - What is Digital Preservation Metadata
  - Hands-on Exercise
  - Case Study: eJournals (1)
- **Preservation Metadata in Practice**
  - Workflow Issues
  - Tools and Standards
  - PREMIS Data Dictionary
  - Overview
    - Hands-on Exercise
    - Implementation Issues
  - Case Study: eJournals (2)

Creating Metadata Profiles

3. Which standard do we use for which metadata?

- Has your organization adopted a metadata standard that supports digital preservation?
- Has your organization adopted a metadata container format?
- Are you adapting community tools for metadata processing?
- Which use cases are supported by which standard?
- Do you want to support duplicated information?

Question 3: Which standard do we use for which metadata?

- **For eJournals:**

- **METS:**
  - Structural relationships between files
  - File location
  - Digital library system identifiers
  - Basic technical metadata
  - Bundling up remaining metadata

Question 3: Which standard do we use for which metadata?

- **For eJournals:**

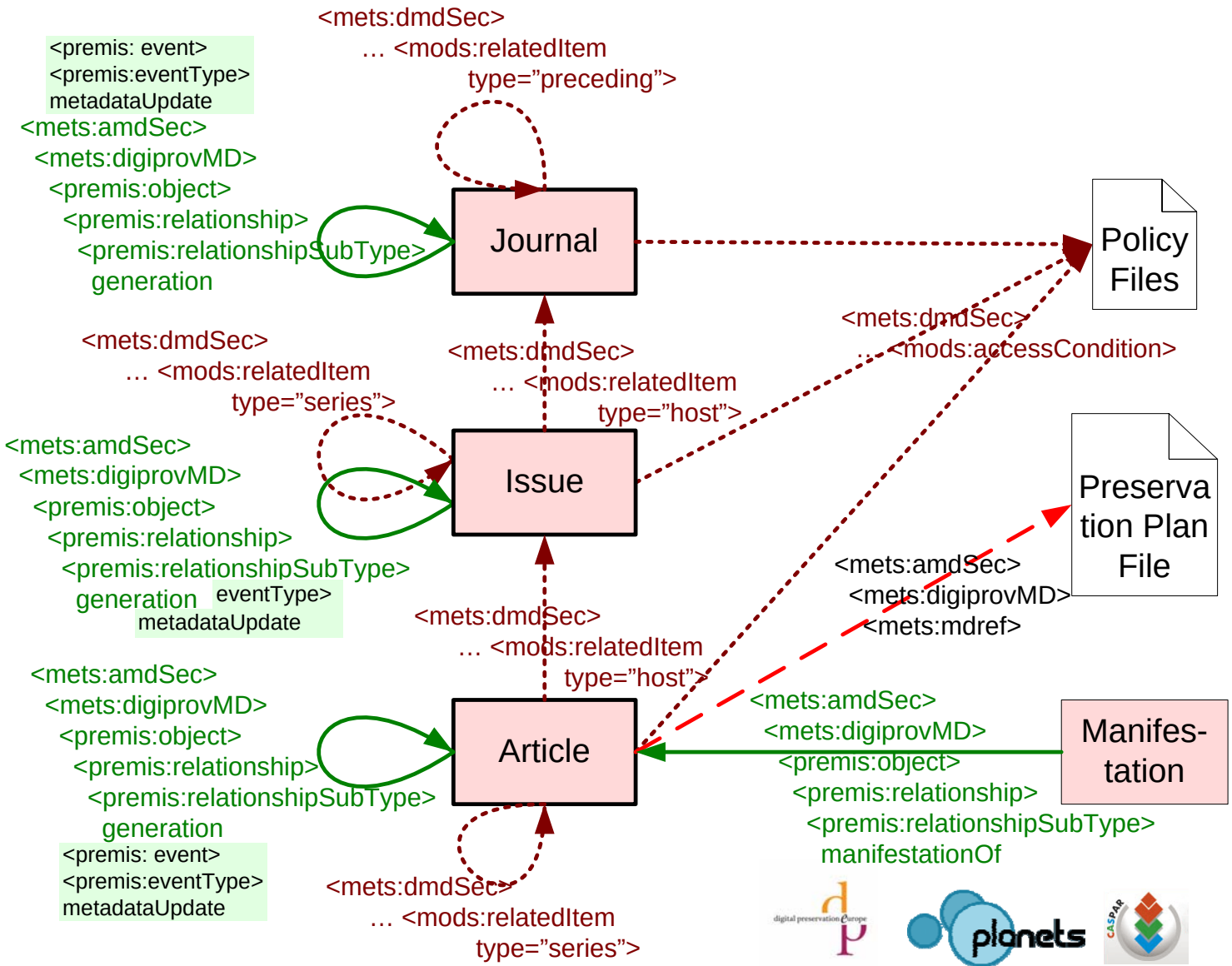
- **MODS:**
  - Descriptive metadata
  - Non-actionable, descriptive rights
  - Relationships between intellectual entities which describe structural or other semantics
  - Identifiers of intellectual entities
  - Provenance information of the record

Question 3: Which standard do we use for which metadata?

- **For eJournals:**

- **PREMIS:**
  - Events (provenance of the content)
  - Agents
  - Basic technical metadata
  - Specific technical metadata
  - Identifiers for AIP generations

Example Diagram



Questions about Repository Design

- Are you creating preservation metadata automatically or manually through user submission or input?
- What will it take to make new or legacy digital objects ready for long-term preservation?
- If you use a third-party repository application, does it accommodate your metadata needs?
- Does the system save metadata in archival storage along with content objects, as well as keeping a working copy to support repository operations?
- Will the repository be able to export standards-conformant metadata according to published XML schema?

- **Thanks**

"The PREMIS Data Dictionary: Information you need to know for preserving digital documents" please use the following license: