TIMBUS

TIMELESS BUSINESS

Sponsored by the European Commission Directorate for Information Society

Digital Preservation Metadata - The PREMIS Data Dictionary

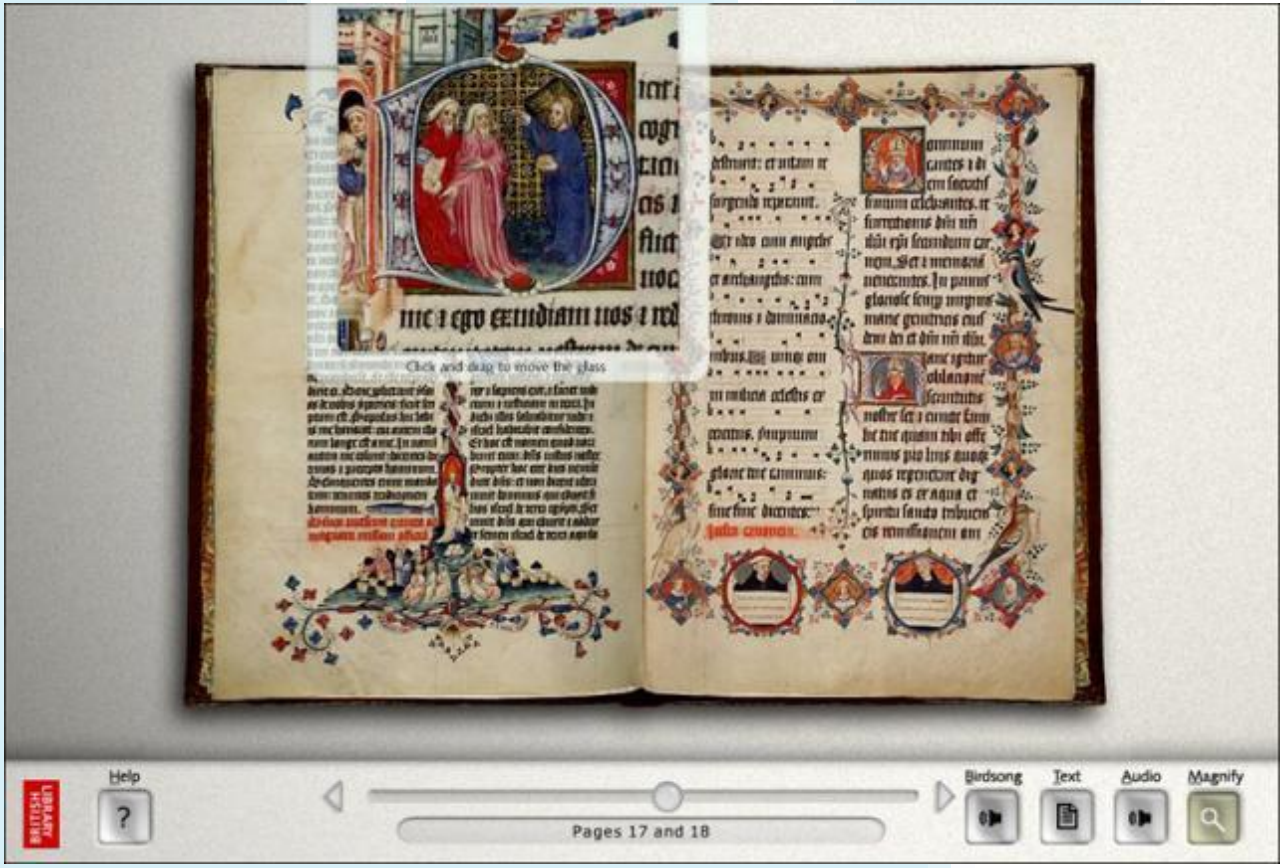
Angela Dappert
angela@dpconline.org

What is Digital Preservation Metadata?

Metadata = data about data

Digital Preservation Metadata =
metadata that is essential to ensure long-term accessibility of digital resources

Domain



What Digital Preservation Metadata to store?

- A best guess on the future
  - little experience validating the longevity of digital objects
  - uncertain future technical possibilities
  - uncertain future legal framework
- 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)

The PREMIS Data Dictionary

Information you need to know for preserving digital documents

Preservation Metadata: Implementation Strategies

The PREMIS Data Model includes

- **Entities: "things" relevant to digital preservation that are described by preservation metadata**
- **Relationships between Entities**
- **Properties of Entities (semantic units)**

- **Object**
  - **Intellectual Entity**
  - **Representation**
  - **File**
  - **Bitstream**
- **Event**
- **Right**
- **Agent**

The process properties: context

Provenance information: logs, business motivations, preservation objectives, design decisions

Activities

- Data Dictionary (PREMIS 2.1)
  - http://www.loc.gov/standards/premis/v2/premis-2-1.pdf
- PREMIS Implementors' Group Forum (pig@loc.gov)
  - email message to listserv@loc.gov subscribe pig your name

Fore Example: Object Entity semantic units

1.1 object Identifier

1.2 object Category

1.3 preservation Level

1.4 significant Properties

1.6 original Name

1.7 storage

1.9 signature Information

1.8 environment
1.8.1 environmentCharacteristic
1.8.2 environmentPurpose
1.8.3 environmentNote
1.8.4 dependency
1.8.5 software
1.8.6 hardware

1.5 objectCharacteristics
1.5.1 compositionLevel
1.5.2 fixity
1.5.3 size
1.5.4 format
1.5.5 creatingApplication
1.5.6 inhibitors

1.10 relationship
1.11 linkingEventIdentifier
1.13 linkingRightsStatementIdentifier

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.		

Scope

What PREMIS DD is:

- Common data model for organizing/thinking about preservation metadata
- Implementable
- Standard for exchanging information packages between repositories
- Technically neutral
- Core metadata

Scope

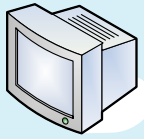
What PREMIS DD is not:

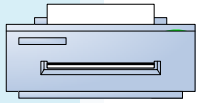
- Out-of-the-box solution
- All needed metadata
- Lifecycle management of objects outside repository
- Rights management

Why do we need new forms of preservation metadata?

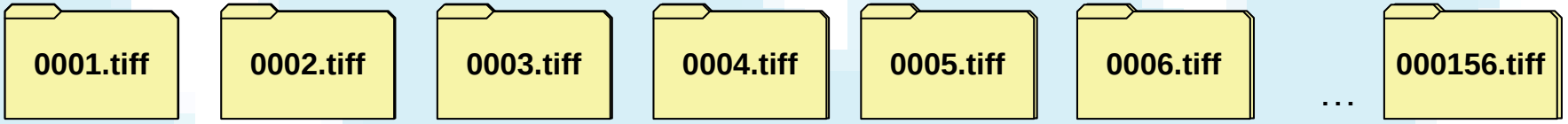
Technology Dependence

digital





Complex environments



No direct access

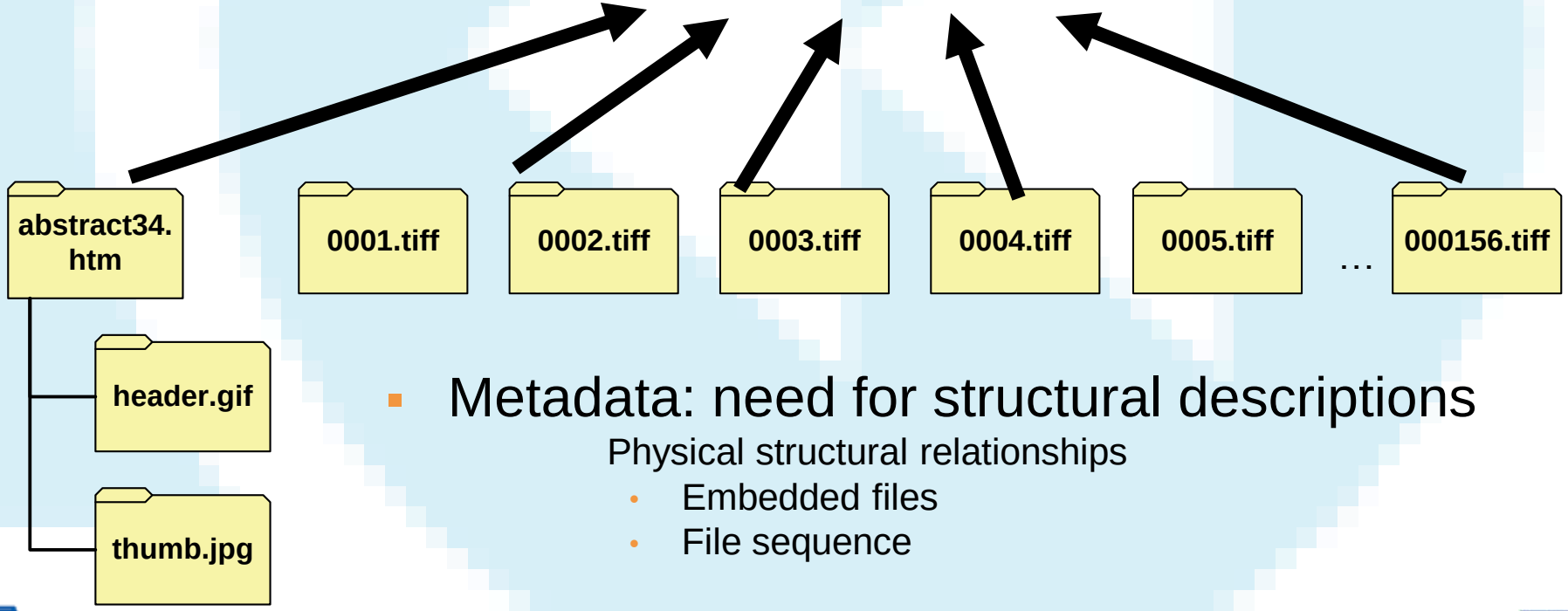
- Not self-descriptive
- Complex formats

Technology Dependence

Metadata:

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

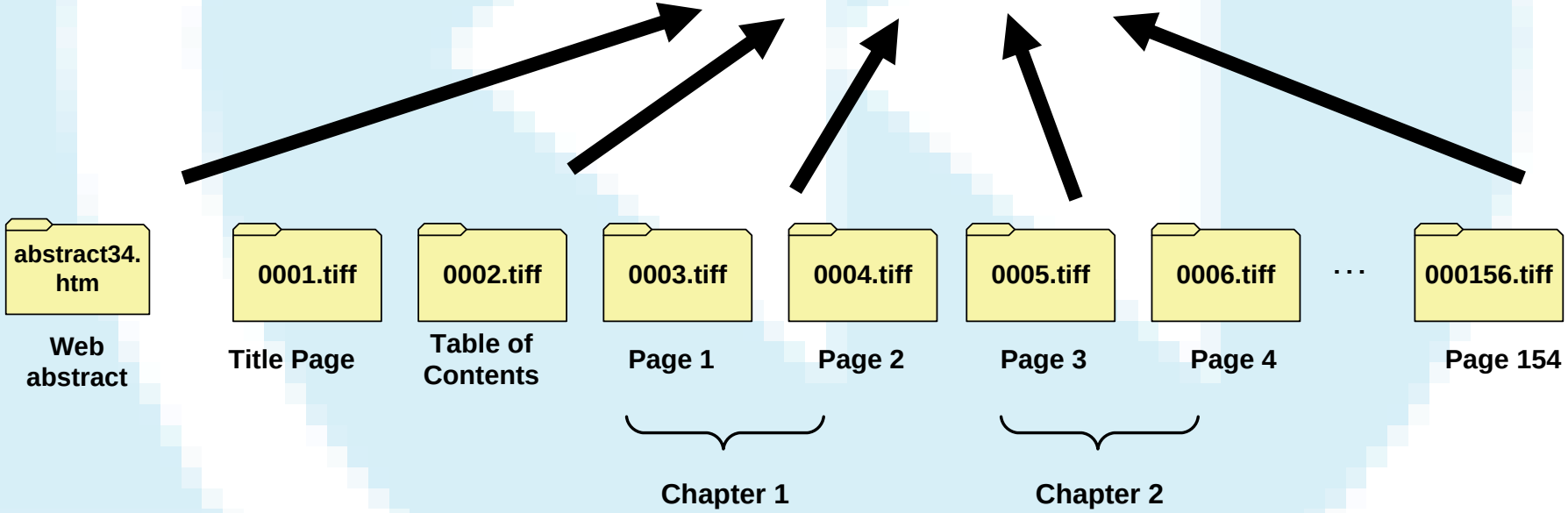
Complex Structures



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

Complex Structures

- Metadata: need for structural descriptions
  - Logical structural relationships



Supporting New Features

- Metadata:
  Semantic Information for the designated community

Obsolescence

Action:

- **Frequent, pre-emptive preservation actions (migration, emulation)**

Metadata:

- **Provenance metadata:**
  - **history of all actions performed on the resource**
    - **events**
    - **changes and decisions**
    - **agents (decision maker + tools used)**
    - **dates**
  - **history of custodianship**
- **Business rules guiding preservation actions**

Obsolescence

Action:

- **Preservation actions during copyright period**

Metadata:

- **Preservation action rights information**

Obsolescence

Action:

- **Preservation actions resulting in potential loss of object characteristics**

Metadata:

- **Significant characteristics = business requirement**
- **Technical and content characteristics of objects before and after preservation actions**

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)**

Metadata:

- **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)**

Metadata:

- **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

Metadata:

- event information for format identification and validation events (= provenance)
- structural metadata

Mutability

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

Action:

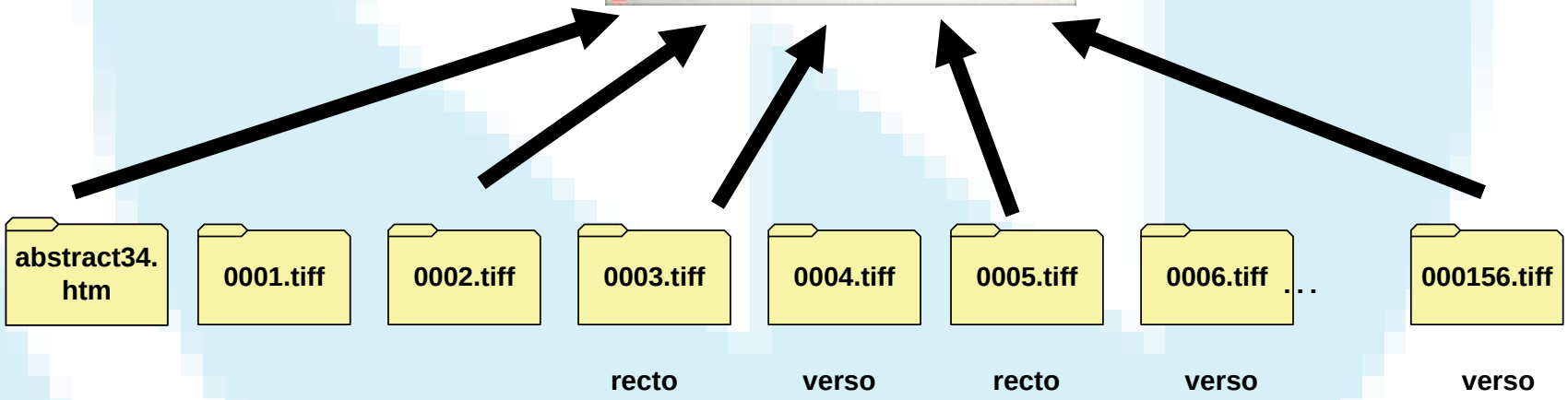
- **Procedural:**
  - virus protection
  - firewalls
  - tight authentication
  - intrusion detection
  - immediate attention to security alerts
- **Technical:**
  - Replication
  - digital signatures

Metadata:

- **Provenance metadata**
- **Digital signatures**
- **Access rights**

Context Descriptions

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



Preservation Pyramid
(from Priscilla Caplan)