DELOS

Network of Excellence on Digital Libraries

Categories, uses and challenges of metadata and process documentation

DELOS International Summer School 2005
INRIA, Sophia Antipolis, France
5-11 June 2005

Michael Day
UKOLN, University of Bath & Digital Curation Centre
http://www.ukoln.ac.uk/

Information Society

DELOS
Network of Excellence on Digital Libraries

Session outline

- What is metadata and how can it support digital preservation?
- Discussion
- Introduction to the PREMIS Data Dictionary and other selected initiatives
- Practical session - Using the PREMIS Data Dictionary
- Summary and concluding discussion

Information Society

DELOS

Network of Excellence on Digital Libraries

Preservation metadata - an introduction

Information Society

Defining metadata (1)

- **Some definitions:**
  - **Literally, "data about data"**
    - **Defines the basic concept, but is (perhaps) not very meaningful**
    - **Refers to everything and nothing (Duff, 2004)**
  - **"Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage" information objects (NISO, 2004)**

Defining metadata (2)

- **Metadata is now typically defined by *function***
  - **For example, the popular categorisation promoted by the Library of Congress:**
    - **Descriptive metadata**
    - **Structural metadata**
    - **Administrative metadata**
  - **But note: potential overlap between categories, the need to consider extrinsic qualities like "context"**

DELOS
Network of Excellence on Digital Libraries

Defining metadata (3)

- The importance of metadata has now been realised:
  - … "is recognised as a critically important, and yet increasingly problematic and complex concept with relevance for information objects as they move through time and space" -- Gilliland-Swetland (2004)
- Metadata needs to be 'understood' by both machines and people

Information Society

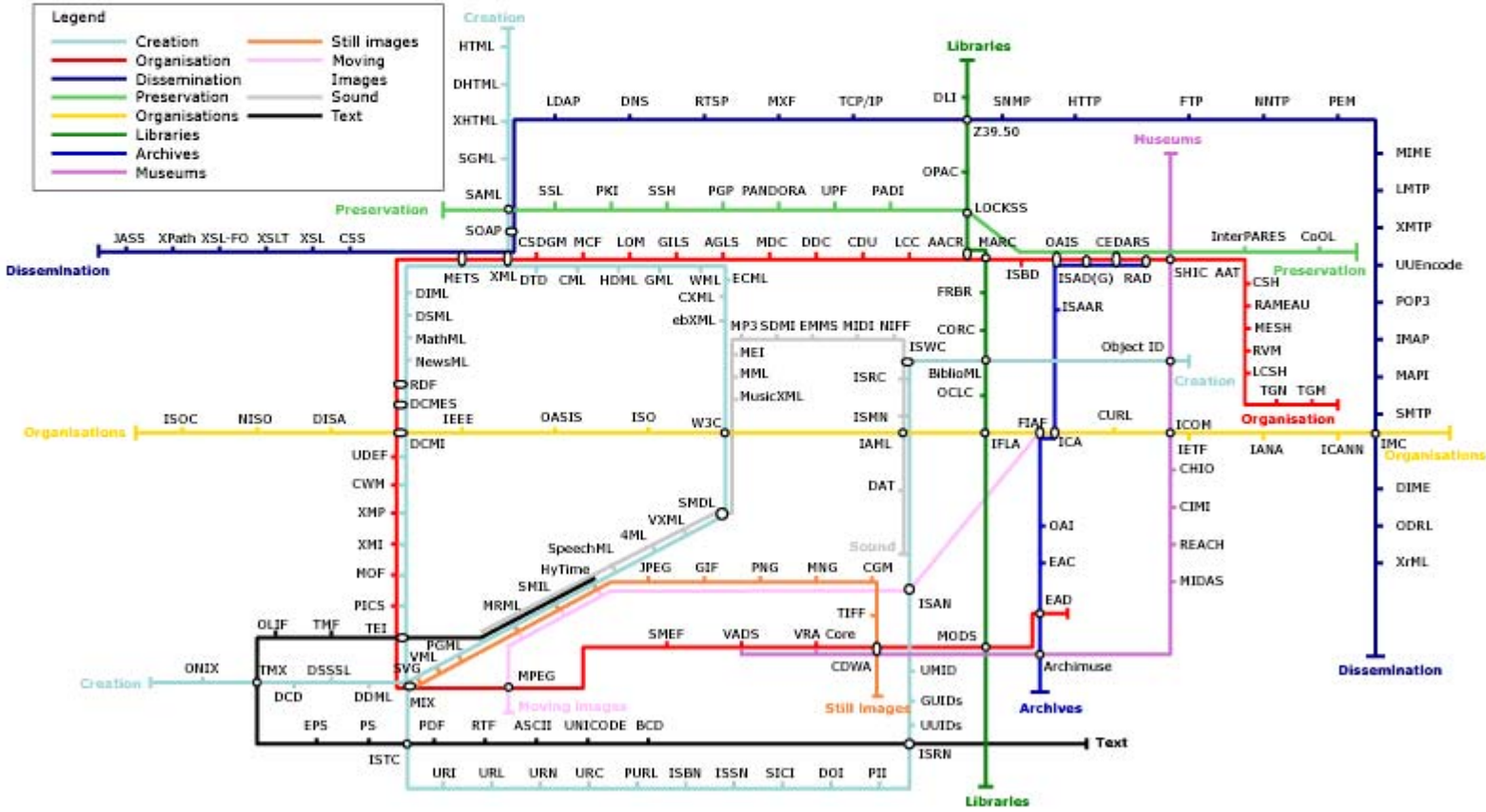
Defining metdata (4)

- **But a large (and growing) number of initiatives, formats, schemas, etc.**
  - **See James Turner's MetaMap for one attempt to visualise the metadata information space: http://mapageweb.umontreal.ca/turner/meta/english/**

MetaMap

[© 2004 James M. Turner, Véronique Moal, & Julie Desnoyers]

Position the mouse over an acronym to see what it stands for in a popup window.
Click on the acronym to see its definition and a link to its official site.



Preservation metadata (1)

- **Definitions:**
  - **All of the various types of data that allow the re-creation and interpretation of the structure and content of digital data over time (Ludäsher, Marciano and Moore, 2001)**
  - **"… the information a repository uses to support the digital preservation process" -- PREMIS working group (2005)**

Preservation metadata (2)

- ***All*** **digital preservation strategies depend, to some extent, upon the creation, capture and maintenance of appropriate metadata**
- **"Preserving the right metadata is key to preserving digital objects" -- ERPANET Briefing Paper (Duff, Hofman & Troemel, 2003)**

DELOS
Network of Excellence on Digital Libraries

Preservation metadata (3)

- **Preservation metadata fulfil a range of different roles, e.g.:**
  - **"… metadata accompanies and makes reference to each digital object and provides associated descriptive, structural, administrative, rights management, and other kinds of information" (Lynch, 1999)**
  - **Spans the categories of administrative, structural, descriptive and technical metadata**

Information Society

DELOS
Network of Excellence on Digital Libraries

Preservation metadata (4)

- **Metadata is key to the understanding and reuse of digital information, e.g.:**
  - **"… it is impossible to conduct a correct analysis of a data set without knowing how the data was cleaned, calibrated, what parameters were used in the process, etc." -- Deelman, *et al*. (2004)**
  - **Growing emphasis on open access to research data (OECD working group)**
  - **The 'data deluge'**

Information Society

Preservation metadata (5)

- **Wide range of relevant initiatives:**
  - **Cultural heritage information**
    - **Supporting repository functions (e.g., PREMIS Data Dictionary, METS, …)**
    - **Digital imaging (NISO Z39.87, …)**
  - **Archives (RKMS, VERS, InterPARES, …)**
  - **Scientific data**
  - **Multimedia (MPEG-7, MPEG-21, …)**

Preservation metadata (6)

- **Current position:**
  - **Early initiatives tended to be theoretical in nature (e.g., metadata frameworks); current ones have a far more practical focus**
  - **Some consensus in cultural heritage domain on the *types* of metadata required**
    - **Influence of the Reference Model for an Open Archival Information System (OAIS)**

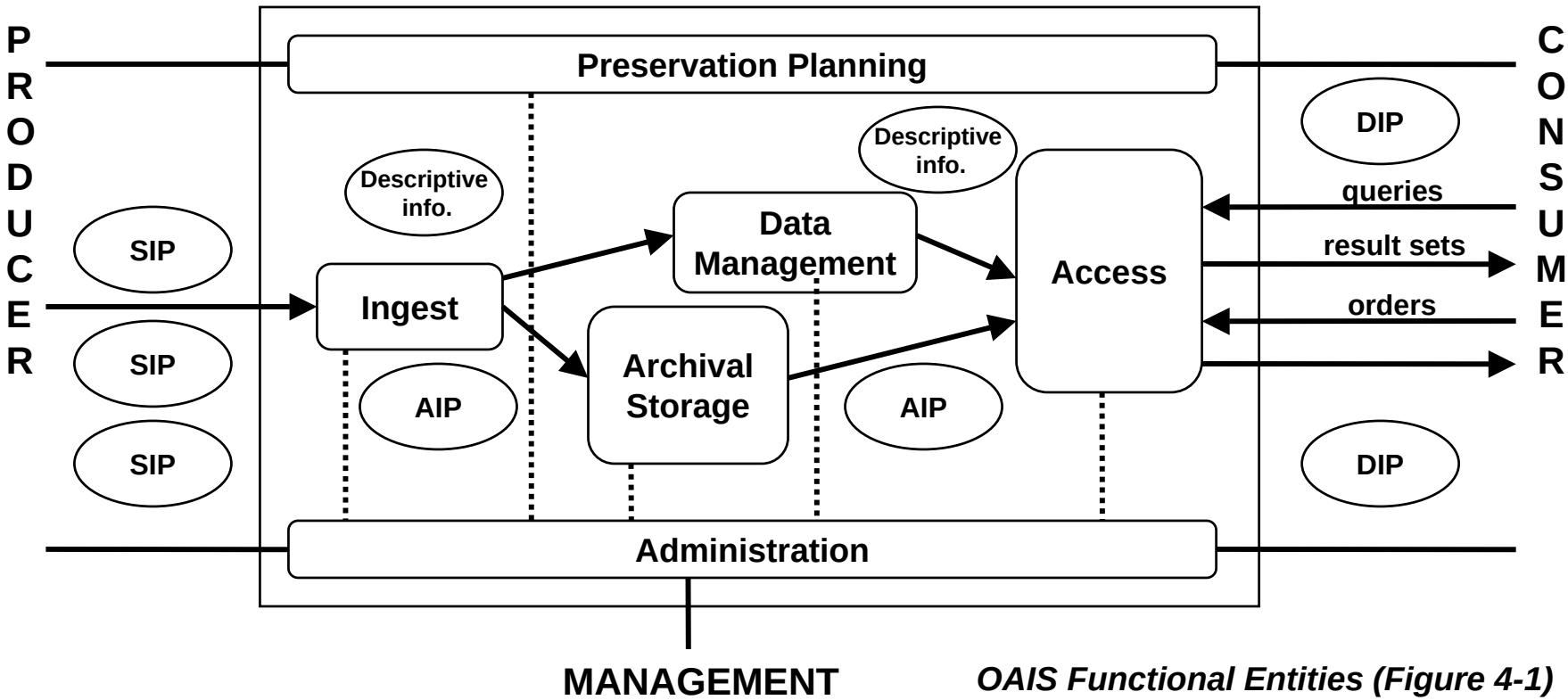
DELOS
Network of Excellence on Digital Libraries

The OAIS model

- ISO 14721:2003
- A reference model - establishes a common framework of terms and concepts
- Identifies the basic functions of an OAIS:
  - Ingest, Data Management, Archival Storage, Administration, Access, Preservation Planning
- Defines an information model

Information Society

OAIS functional model



OAIS Functional Entities (Figure 4-1)

OAIS information model (1)

- **Information Object (basic concept):**
  - **Data Object (bit-stream)**
  - **Representation Information**
    - **Information that permits "the full interpretation of Data Object into meaningful information"**
    - **Technical and structural metadata**

OAIS information model (2)

- **Information Object Classes:**
  - **Content Information**
  - **Preservation Description Information (PDI)**
  - **Packaging Information**
  - **Descriptive Information**

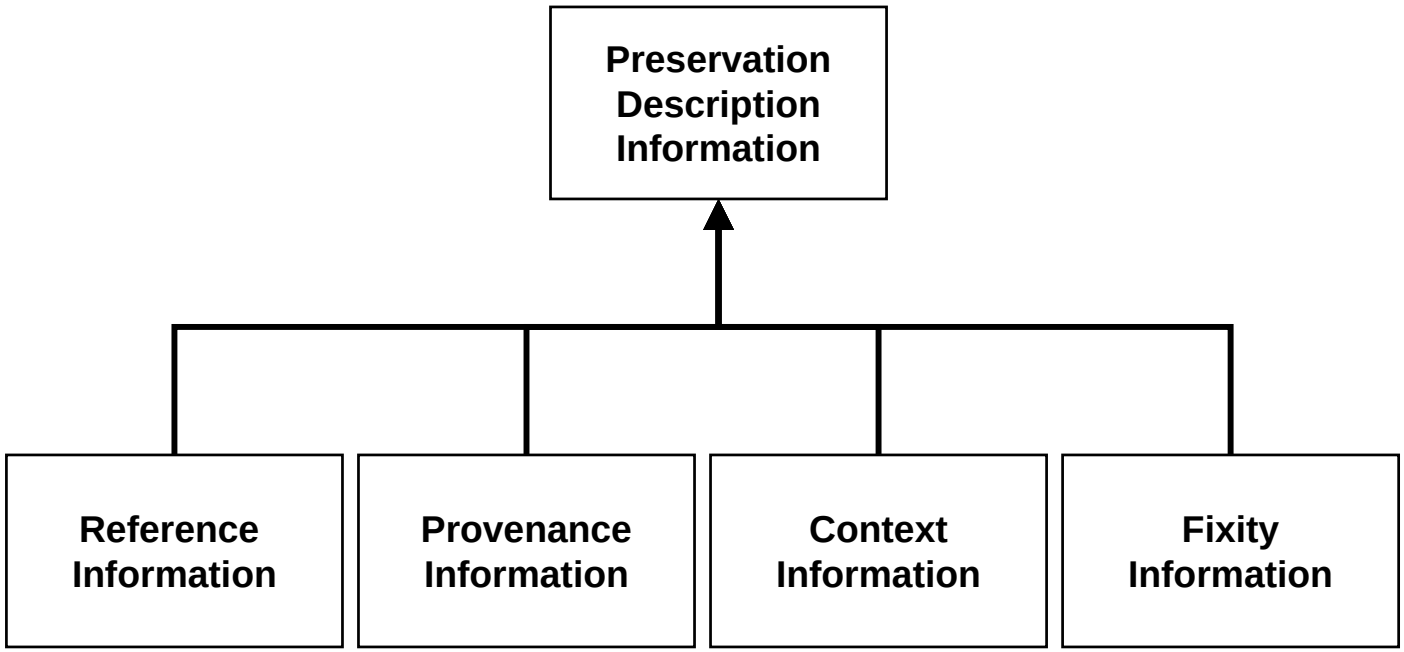
OAIS information model (3)

- **Information Package:**
  - Container that encapsulates Content Information and PDI
  - Packages for submission (SIP), archival storage (AIP) and dissemination (DIP)
  - AIP = "... a concise way of referring to a set of information that has, in principle, all of the qualities needed for permanent, or indefinite, Long Term Preservation of a designated Information Object"

OAIS information model (4)

- **Archival Information Package (AIP):**
  - **Content Information**
    - **Original target of preservation**
    - **Information Object (Data Object & Representation Information)**
  - **Preservation Description Information (PDI)**
    - **Other information (metadata) "which will allow the understanding of the Content Information over an indefinite period of time"**

OAIS information model (5)



PDI Preservation Description Information (Figure 4-16)

DELOS
Network of Excellence on Digital Libraries

Types of PDI (1)

- **Reference Information:**
  - **Reflects the need for objects to be identified and located definitively over time**
  - **A role for descriptive metadata (?)**
  - **Unique identification:**
    - **Within a repository (essential to the management of objects)**
    - **Within the wider environment (unique identifiers)**

DELOS
Network of Excellence on Digital Libraries

Types of PDI (2)

- **Provenance Information:**
  - **Context Information that documents the history of the Content Information**
  - **A key part of the *integrity* of objects is being able to trace their origin and chain of custody**
  - **A principle of archives profession**

Information Society

DELOS
Network of Excellence on Digital Libraries

Types of PDI (3)

- **Context Information:**
  - **Documenting the relationships of the Content Information to its environment**
  - **Understanding of objects cannot properly be interpreted without some understanding of its wider environment, e.g.:**
    - **Technical dependencies on hardware and software**
    - **Linking scientific data to its experimental context**

Types of PDI (4)

- **Fixity Information:**
  - **Reflects the need for the users of digital resources to have confidence that they are what they claim to be and that their integrity have not been compromised**
  - **Integrity checks at the level of Content Data Objects (checksums, etc.)**
  - **Wider issue of provenance, custody, and trust**

DELOS

Network of Excellence on Digital Libraries

Discussion

Information Society

DELOS
Network of Excellence on Digital Libraries

Discussion questions (1)

- What are the most important types of preservation metadata?
- What factors should repositories take into account when defining what preservation metadata they need?
- Are there any major things missing from the OAIS categories?

Information Society

Discussion questions (2)

- What tools may be needed?
- Where should metadata be stored?
- What about interoperability (e.g. object or metadata exchange)?

DELOS

NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

An overview of some metadata initiatives related to preservation

Information Society

Archives

- **Recordkeeping metadata**
  - **Business Acceptable Communications (BAC) model developed by the Pittsburgh Project (1995)**
  - **Australian Recordkeeping Metadata Schema (RKMS)**
  - **Individual standards developed, e.g. by the UK National Archives, the National Archives of Australia, the Public Record Office Victoria, etc.**

DELOS
Network of Excellence on Digital Libraries

Digitisation initiatives

- NISO Z39.87 Technical Metadata for Digital Still Images (draft, 2001)
- Metadata Encoding & Transmission Standard (METS)
  - Maintained by the Library of Congress
  - XML container for different types of metadata: descriptive, administrative, and structural
  - Potential as OAIS Information Package

DELOS
Network of Excellence on Digital Libraries

Research libraries

- An urgent practical response to digital initiatives and the growing amount of digital content needing management:
  - National Library of Australia (1999)
  - Harvard University Library
  - National Library of New Zealand (2003)
- Research projects
  - UK Cedars project outline specification (2000)
  - NEDLIB project (2000)

Information Society

DELOS
Network of Excellence on Digital Libraries

International activity

- **Metadata Framework Working Group**
  - **Sponsored by OCLC and RLG**
  - **Preservation Metadata Framework (2002)**
    - **built upon OAIS model and the work of earlier initiatives**
  - **Framework was a set of recommendations, not a specification for implementation**
- **PREMIS Working Group**

Information Society

DELOS

Network of Excellence on Digital Libraries

The PREMIS Data Dictionary for Preservation Metadata

Information Society

PREMIS Working Group (1)

- **PREMIS WG = Preservation Metadata: Implementation Strategies**
  - **Sponsored by OCLC and RLG**
  - **Established 2003**
  - **International working group and advisory committee (practical focus)**
    - **Members from the US, the UK, the Netherlands, Germany, Australia and New Zealand**
  - **Chaired by Priscilla Caplan and Rebecca Guenther**

DELOS
Network of Excellence on Digital Libraries

PREMIS Working Group (2)

- **Main objectives:**
  - **A 'core' set of preservation metadata elements (Data Dictionary)**
  - **Strategies for encoding, packaging, storing, managing, and exchanging metadata**
- **Outputs:**
  - **Implementation Survey report (Sept. 2004)**
  - **PREMIS Data Dictionary (May 2005)**

Information Society

PREMIS review (1)

- **Implementing Preservation Repositories for Digital Materials**
  - **Review of current practice within cultural heritage organisations**
    - **Based on responses to questionnaire together with follow-up interviews**
    - **Questions about business plans, policies, preservation strategies, as well as metadata**

PREMIS review (2)

- **Findings:**
  - **Very little current experience of digital preservation; no knowledge whether the metadata collected will be adequate**
  - **The OAIS model has informed the implementation of many repositories**
  - **METS was the most commonly-used scheme for non-descriptive metadata**
  - **Metadata is stored *both* in databases and together with content data objects**

PREMIS review (3)

- Trends identified:
  - Redundant storage of metadata both within databases (for ease of use) and encapsulated with data objects (self-documenting)
  - METS is commonly used for the packaging of different metadata
  - OAIS is just the starting point
  - The retention of the original versions of objects to reduce risks
  - The use of multiple preservation strategies

[http://www.oclc.org/research/projects/pmwg/]

Data Dictionary for Preservation Metadata

Contents:

DELOS
Network of Excellence on Digital Libraries

PREMIS data dictionary (1)

- **Background:**
  - **OAIS remains the conceptual foundation (but some differences in terminology)**
  - **The data dictionary is a translation of the OAIS-based 2002 *Framework* into a set of implementable semantic units**
  - **Preservation metadata = "the information a repository uses to support the digital preservation process"**

Information Society

DELOS
Network of Excellence on Digital Libraries

PREMIS data dictionary (2)

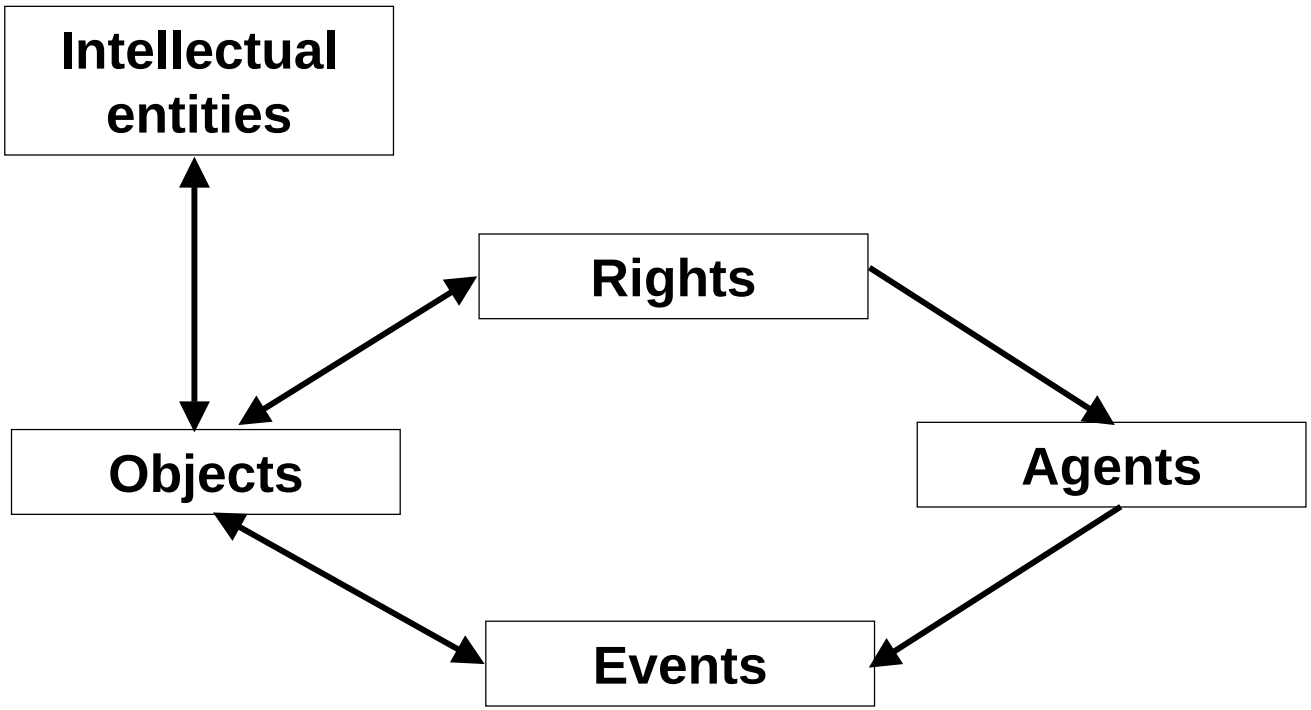
- Defines metadata that supports "maintaining viability, renderability, understandability, authenticity, and identity in a preservation context."
- Core metadata = "things that most working repositories are likely to need to know in order to support digital preservation."
- Recognition of the need for automatic capture of metadata

Information Society

PREMIS data dictionary (3)

- The Data Dictionary is implementation independent, i.e. does not define how it should be stored
- Based on simple data model that defines five types of entities

PREMIS data model (1)



PREMIS data model (2)

- *Entities*:
  - Digital Object, Intellectual Entity, Event, Agent, & Rights
- *Relationships* are statements of association between instances of entities
- *Semantic Units* are the properties of an entity, and have values

PREMIS data model (3)

- *Digital Object* = a discrete unit of information
  - Files = named and ordered sequence of bytes known by an operating system
  - Bitstream = a set of bits embedded within a file
  - Representation = the set of files needed for a "complete and reasonable" rendering of an Intellectual Entity

PREMIS data model (4)

- *Intellectual Entity* = a coherent set of content that can be viewed as a single unit
- *Event* = an action involving at least one Object or Agent known to the repository
  - Documents actions that modify Digital Objects, records validity checks, etc.
  - Objects can be associated with any number of events

PREMIS data model (5)

- *Agent* = persons, organisations, or programs associated with preservation events
  - Not the main focus of the data dictionary
- *Rights Statements* = assertions of rights pertaining to Objects or Agents
  - WG concentrated on rights and permissions associated with preservation activities

PREMIS data model (6)

- ***Relationships:***
  - **Relationships between Objects:**
    - **Structural relationships, e.g. how files combine to make up an Intellectual Entity**
    - **Derivation relationships, e.g. resulting from format transformations or replications**
    - **Dependency relationships, e.g. when Objects depend on others, e.g. fonts, DTDs, etc.**
  - **1:1 principle**

Data Dictionary, v 1.0 (1)

- **Defines semantic units for Objects, Events, Agents and Rights**
  - **Object: objectIdentifier, preservationLevel, objectCategory, objectCharacteristics (format, significant properties, etc.), creatingApplication, storageMedium, environment (dependencies, hardware and software details, etc), relationship, …**

DELOS
Network of Excellence on Digital Libraries

Data Dictionary v 1.0 (2)

- Event: eventIdentifier, eventType (from a controlled list, e.g. ingestion, migration, normalization), eventDateTime, eventDetail, eventOutcomeInformation, linkingAgentIdentifier, …
- Agent: agentIdentifier, agentName, agentType, …
- Rights: permissionStatement, …

Information Society

Limits to scope (1)

- Does not focus on descriptive metadata
  - Domain specific and dealt with by many other schemes
- Does not define the characteristics of Agents
- Does not directly consider rights and permissions not directly associated with preservation actions, e.g. access or reuse

DELOS
Network of Excellence on Digital Libraries

Limits to scope (2)

- Does not deal with technical metadata for all different types of digital file (left to format experts)
- Does not deal with the detailed documentation of media or hardware (left to specialists)
- Does not consider in detail the business rules of a repository, e.g. roles, policies, and strategies (but this could be added to data model)

Issues (1)

- The PREMIS Data Dictionary is an important contribution to the ongoing development of preservation metadata
- It is, however, implementation independent
  - Provides definition of semantics and a suggested XML binding
- Maintenance Agency (Library of Congress):
  - http://www.loc.gov/standards/premis/schemas.html

Issues (2)

- Conformance
  - Non-PREMIS elements not conflict with or overlap with PREMIS semantic units
  - Need for more harmonisation
- The exchange of Objects
  - Mandatory metadata needs to be able to be extracted and packaged with the object
- The use of controlled vocabularies

DELOS

Network of Excellence on Digital Libraries

Using the PREMIS Data Dictionary for Preservation Metadata (practical session)

Information Society

Practical exercise (1)

- Go quickly through the semantic units for the Object Entity (pp. 2-5 - 2-53 only)
- Identify whether they are intrinsic to the object or whether the metadata content needs to be created or captured by the repository
  - Examples: preservationLevel can only be decided by the repository; the formatName is somehow directly dependent on the object itself

Practical exercise (2)

- Process: pass the Data Dictionary around the group so that you all take it in turns to lead the discussion of a semantic unit
- Not exact process, do not discuss each semantic unit for too long, do not get too bogged down in detail
- The object is to identify what PREMIS metadata *may* be able to be captured automatically by analysis of objects on ingest

Practical exercise (3)

- This should result in two lists:
  - "Intrinsic" metadata that may be able to be captured
  - Metadata that needs to be generated by the repository (or its processes)
- Quick report back at end from groups on the first list only
- Objective is to get to know the detail of the PREMIS Data Dictionary a little better
- Hopefully it will be fun!

Other issues

Costs of metadata

- **Balance risks with costs:**
  - **There is a perception that metadata creation and maintenance will be expensive**
  - **But costs associated with data recovery are not trivial**
- **Avoid imposing unnecessary costs:**
  - **Avoid large schemas**
  - **Need to identify the *right* metadata (importance of 'core metadata')**

Metadata capture

- Need to reduce amount of metadata created 'by hand'
- Capture that metadata that already exists in other forms
- Develop tools to automatically capture some metadata (e.g. technical metadata)
- Need for event metadata to be captured at creation, ingest, migration, and at other appropriate points in the object life-cycle

Interoperability

- **Interoperability is important:**
  - **To support the reuse of existing metadata**
  - **To support the exchange of digital objects between repositories**
- **A role for centralised repositories for file format information (e.g. PRONOM, GDFR) and metadata schemas?**

DELOS

Network of Excellence on Digital Libraries

Summing up

Information Society

DELOS
Network of Excellence on Digital Libraries

Summing up

- Metadata is perceived to be essential for the long-term management (preservation) of digital objects - think about the metadata required in every session of this summer school
- There is now some consensus on what metadata might be required to support preservation processes (e.g., OAIS model, PREMIS Data Dictionary)
- Still little experience with the practical implementation of preservation metadata

Information Society

Key links:

- **PREMIS Data Dictionary for Preservation Metadata: http://www.oclc.org/research/projects/pmwg/**
- **ERPANET Training Seminar on "Metadata in Digital Preservation" (Marburg, 2003): http://www.erpanet.org/**
- **OAIS Reference Model: http://www.ccsds.org/documents/650x0b1.pdf**

DELOS
Network of Excellence on Digital Libraries

Acknowledgements

UKOLN is funded by the Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils, as well as by project funding from the JISC, the European Union, and other sources. UKOLN also receives support from the University of Bath, where it is based.

http://www.ukoln.ac.uk/

The *Digital Curation Centre* is funded by the JISC and the UK Research Councils' e-Science Core Programme.

http://www.dcc.ac.uk/

Information Society