Other types of metadata - METS, PREMIS, …

Michael Day
Digital Curation Centre
UKOLN, University of Bath
m.day@ukoln.ac.uk

Cataloguing Online Resources: an Introduction to Metadata for Librarians, Manchester, 26 April 2006

Session overview

- Metadata for managing and preserving resources
- Archives
- Digitisation initiatives
  - METS
- Preservation metadata
  - The OAIS Information Model
  - PREMIS Data Dictionary

Management and preservation

- Early recognition that metadata was not only useful for resource discovery
- Some examples:
  - Records management and archives
  - Digitisation initiatives
  - Digital preservation

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.

Digitisation initiatives

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

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)
  - *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)

Preservation metadata (2)

- 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

Preservation metadata (3)

- 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'

Preservation metadata (4)

- 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
    - Major influence of the Reference Model for an Open Archival Information System (OAIS)

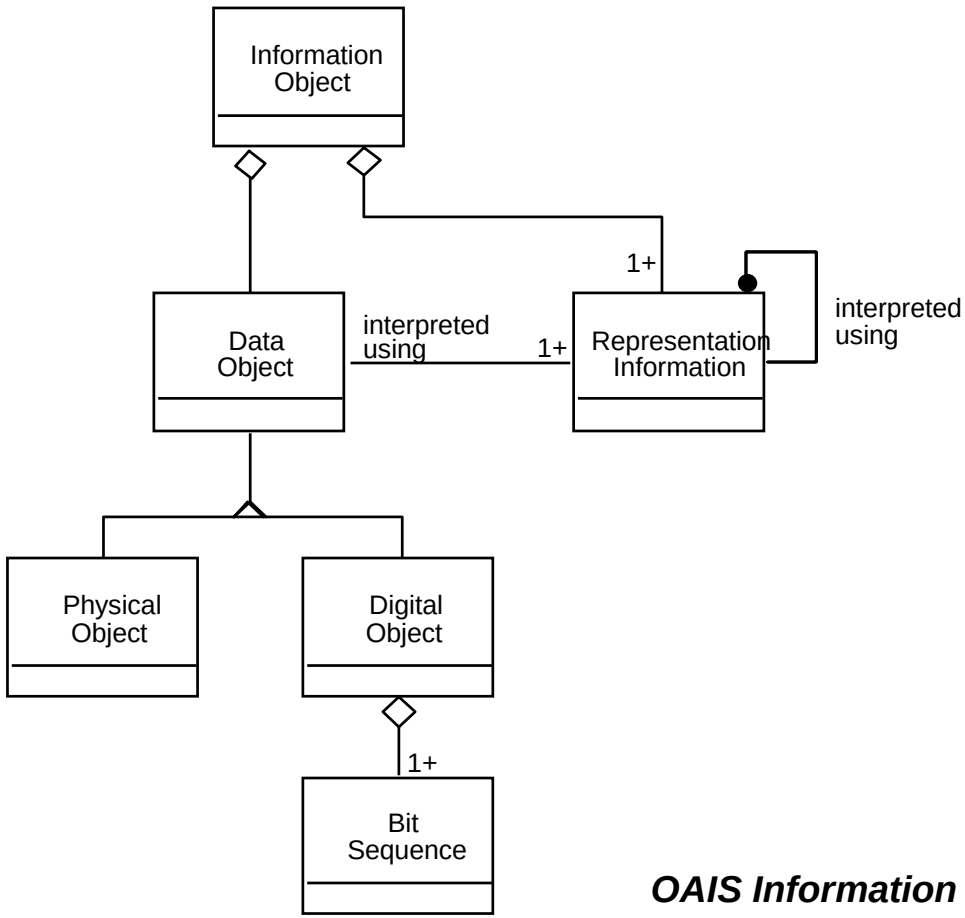
OAIS background

- Reference Model for an Open Archival Information System (OAIS)
  - Development led by the Consultative Committee for Space Data Systems (CCSDS)
  - Issued as CCSDS Recommendation (Blue Book) 650.0-B-1 (January 2002)
  - Also adopted as: ISO 14721:2003
- Defines functional entities and an information model

OAIS Information Model (1)

- Information Object (basic concept):
  - Data Object (bit-stream)
  - Representation Information (permits “the full interpretation of Data Object into meaningful information”)
- Information Object Classes:
  - Content Information
  - Preservation Description Information (PDI)
  - Packaging Information
  - Descriptive Information

OAIS Information Model (2)



OAIS Information Object (Figure 4-10)

OAIS Information Model (3)

- Representation Information:
  - *Any* information required to render, interpret and understand digital data (includes file formats, software, algorithms, standards, semantic information etc.)
  - Representation Information is recursive in nature
  - Essential that Representation Information itself is curated and preserved to maintain access to (render and interpret) digital data
    - e.g. Format registries (GDFR, PRONOM)

OAIS Information Model (5)

- 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 (6)

- 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”
    - A set of Information Objects
    - In part based on categories discussed in CPA/RLG report: *Preserving Digital Information* (1996)

OAIS Information Model (8)

- **Fixity** - supporting data integrity checking mechanisms
- **Reference** - for supporting identification and location over time
- **Context** - documenting the relationship of the Content Information to its environment
- **Provenance** - documents the history of the Content Information

Preservation metadata standards

- Two triggers:
  - An urgent practical response to 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)

OCLC/RLG Metadata Framework

- 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 (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

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)

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:

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"

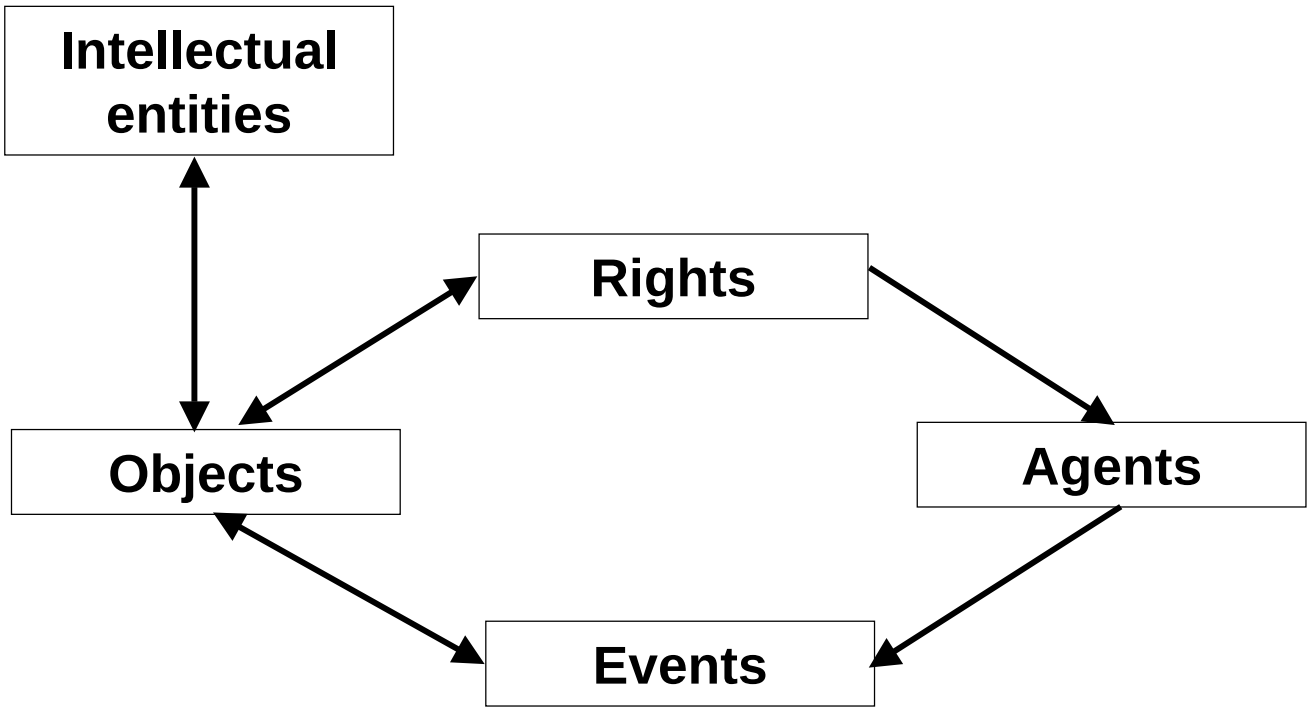
PREMIS data dictionary (2)

- 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

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
- Defines semantic units for Objects, Events, Agents and Rights

PREMIS data model



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

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

METS basics (1)

- Metadata Encoding and Transmission Standard
  - Originated in digitisation projects, i.e. Making of America II
  - An XML-based framework for packaging various types of metadata (and data), including
    - Descriptive - for discovery and retrieval
    - Administrative - enabling managers to administer the object (as part of a collection)
    - Structural Map - how individual components relate to one another

METS basics (2)

- Implemented widely in digital library projects, e.g. Oxford Digital Library
- Supports Interoperability
  - Different metadata can be combined within a METS container, e.g. MODS, MARC in XML, DC in XML, etc.
- Supports the portability of objects
- METS can be seen as a type of Information Package (in OAIS terms), combining data and metadata

Summing up

- Metadata is perceived to be essential for the long-term management and preservation of digital objects
- There is now the beginning of consensus on what particular metadata might be required to support preservation processes (e.g., the OAIS model, PREMIS Data Dictionary) and packaging (e.g. METS)
- There is still little experience with the practical implementation of preservation metadata

Key links:

- PREMIS Data Dictionary for Preservation Metadata: http://www.oclc.org/research/projects/pmwg/
- OAIS Reference Model: http://public.ccsds.org/publications/archive/650x0b1.pdf
- METS: http://www.loc.gov/standards/mets/
- DPC Report on Preservation Metadata: http://www.dpconline.org/
- DCC Digital Curation Manual instalment on Metadata: http://www.dcc.ac.uk/

Other types of metadata - METS, PREMIS, …

Michael Day
Digital Curation Centre
UKOLN, University of Bath
m.day@ukoln.ac.uk

Cataloguing Online Resources: an Introduction to Metadata for Librarians, Manchester, 26 April 2006

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/