Metadata for preservation

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

Chinese-European Workshop on Digital Preservation,
Beijing, China, 14-16 July 2004

Presentation outline

- How can metadata support preservation strategies?
- Current initiatives (brief overview)
- Some key initiatives in more detail:
  - OAIS Reference Model
  - OCLC/RLG Metadata Framework
  - PREMIS working group
- Some issues:
  - Implementation, metadata creation and capture, sustainability, interoperability

Why metadata is useful (1)

- Digital preservation strategies - migration, emulation, technology preservation, etc. - all depend - to some extent - on the creation, capture and maintenance of suitable metadata:
  - "Preserving the right metadata is key to preserving digital objects" (ERPANET Briefing Paper, 2003)
  - "It's all about metadata" (Cedars project manager, ca. 2000)

Why metadata is useful (2)

- Metadata fulfil various roles, e.g.:
  - Within a digital repository, “metadata accompanies and makes reference to each digital object and provides associated descriptive, structural, administrative, rights management, and other kinds of information” (Clifford Lynch, 1999)

Some examples (1)

- Digital libraries
  - National Library of Australia (1999)
  - Cedars project outline specification (2000)
  - NEDLIB project (2000)
  - OCLC/RLG working group metadata framework (2002)
  - National Library of New Zealand (2003)
  - PREMIS working group (2003-)

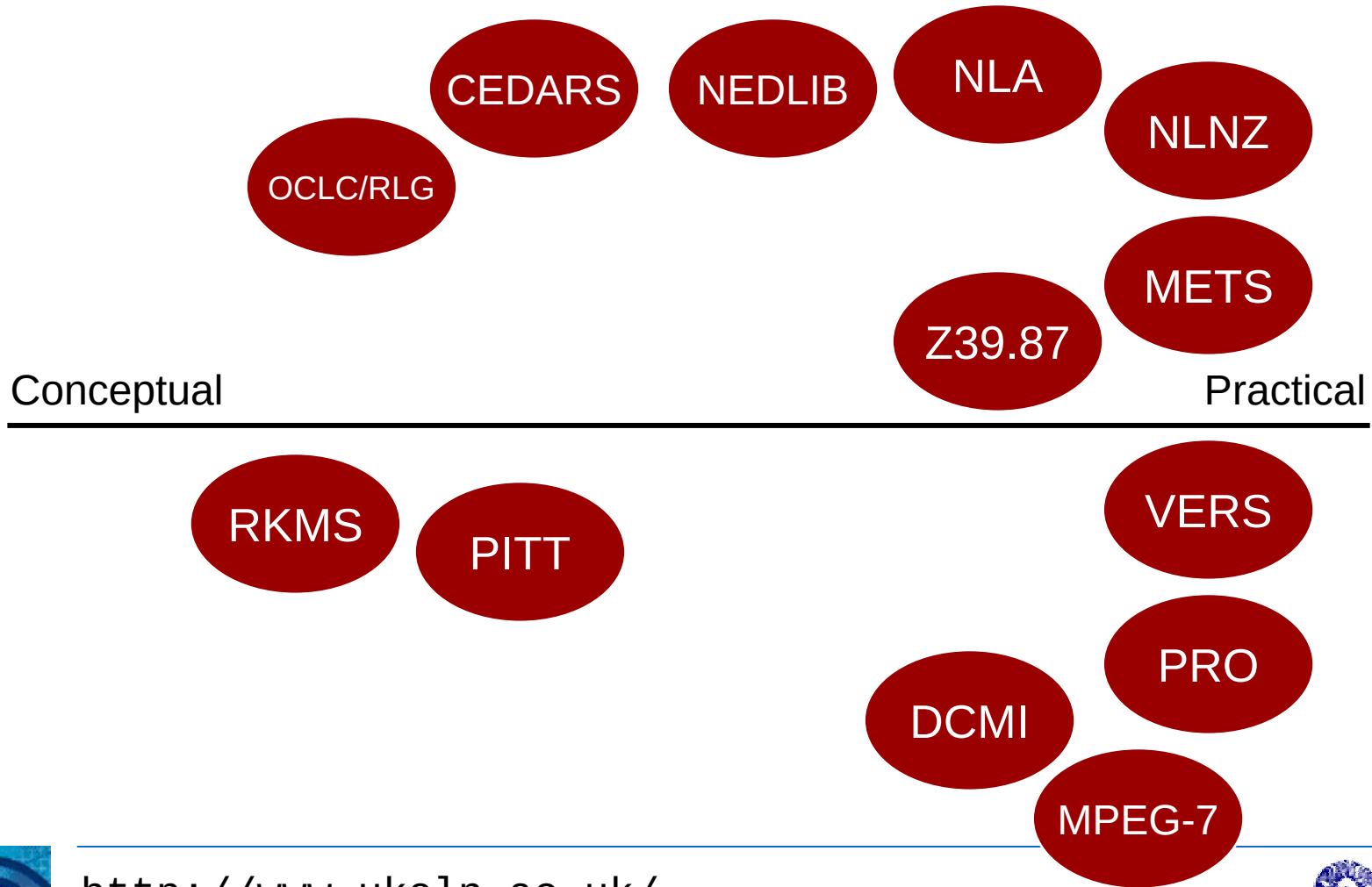
Some examples (2)

- Digitisation
  - NISO Technical Metadata for Digital Still Images (draft, 2001)
  - Metadata Encoding & Transmission Standard (METS)
    - XML container for different types of metadata, descriptive, administrative, structural
    - Supported by Library of Congress

Some examples (3)

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

Draft categorisation (1)



Draft categorisation (2)

- Earliest schemas were largely conceptual in nature:
  - e.g. Pittsburgh BAC model, Cedars outline specification, OCLC/RLG WG
- Gradually moving towards a more practical focus:
  - e.g., VERS, NLNZ, METS, PREMIS
  - Based on XML (DTDs and Schemas)
- But there is an urgent need for this experience to be shared
  - e.g., briefing papers, advice to implementers

The OAIS reference model (1)

The Reference Model for an Open Archival Information System (OAIS):

- ISO 14721:2003
- Establishes a common framework of terms and concepts
- Identifies basic functions of an OAIS:
  - Ingest, Data Management, Archival Storage, Administration, Access, Preservation Planning
- Defines an information model, e.g.:
  - Information Packages
  - Identifies the types of metadata required (but not a schema)

The OAIS reference model (2)

- Information model:
  - 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

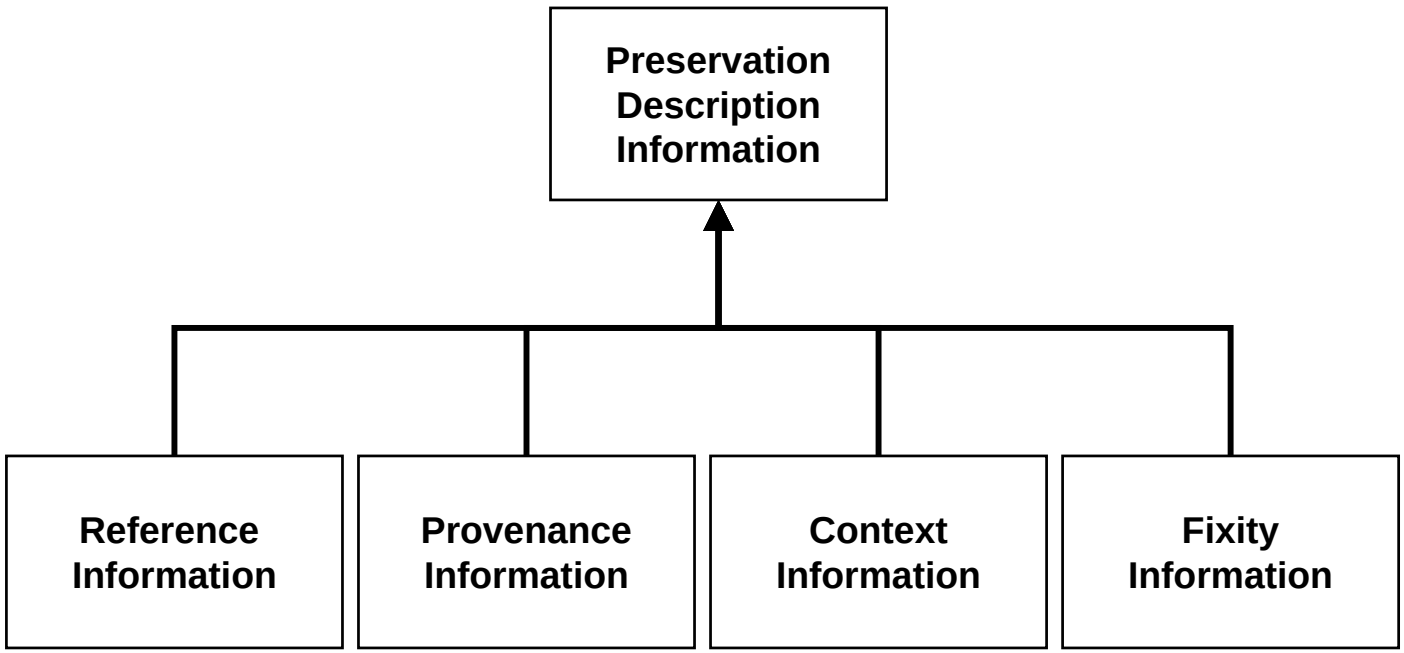
The OAIS reference model (3)

- Information model (continued):
  - 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"

The OAIS reference 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”
    - A set of Information Objects
    - Based on categories discussed in CPA/RLG report: *Preserving Digital Information* (1996)

The OAIS reference model (5)



PDI Preservation Description Information (Figure 4-16)

OCLC/RLG Framework (1)

- Content Information recommendation:
  - The content and all information required to render it
  - OAIS Representation Information - permits “the full interpretation of Data Object into meaningful information”
  - Content Data Object Description, e.g.:
    - Underlying abstract form description
    - Structural type (e.g. MIME type)
    - Technical infrastructure (internal structure)

OCLC/RLG Framework (2)

- Content Information (continued)
  - Content Data Object Description, e.g.:
    - File description (technical specifications)
    - Size
    - Significant properties
  - Environment description
    - Describes the hardware and software environment
    - Operating systems and rendering programs
    - Storage, computational resources and peripherals
    - Available documentation

OCLC/RLG Framework (3)

- Preservation Description Information recommendation:
  - PDI = other information (metadata) “which will allow the understanding of the Content Information over an indefinite period of time” (OAIS Reference Model), e.g.:
    - *Reference*: identifiers (internal and external), basic resource description, existing descriptive metadata
    - *Context*: context of creation, relationships with other data objects

OCLC/RLG Framework (4)

- PDI Recommendation (continued)
  - *Provenance*: event based model, documents an object's origin (creation), existence before ingest, processes enacted at ingest and for maintenance (e.g. migration); also records rights management information
  - *Fixity*: records authenticity procedures

- Framework is a set of recommendations, not a specification for implementation

PREMIS working group (1)

- Working Group on Preservation Metadata - Implementation Strategies
  - Background:
    - Sponsored by OCLC Online Computer Library Center and Research Libraries Group (RLG)
    - WG I (2000-2002) produced state of the art report and metadata framework
    - WG II (PREMIS) focused on implementation

PREMIS working group (2)

- Before WG I
  - Little consensus in digital library world (various projects and initiatives)
  - Awareness of importance of OAIS model, but less understanding of how this should be used
- The PREMIS working group:
  - 2003 - 2004
  - Chairs: Priscilla Caplan and Rebecca Guenther
  - International group from the US, the UK, the Netherlands, Germany, Australia and New Zealand

PREMIS working group (3)

- Aims:
  - Define 'core' set of metadata elements (data dictionary)
  - Evaluate strategies for encoding, storing, managing, and exchanging metadata
- Activities
  - Review WG I framework element by element
  - Focus on high-level, e.g. detailed format-specific metadata out of scope
  - Relationships between digital objects (complex)
  - Survey on metadata requirements of repositories

Issues - implementation

- Focus on implementation is becoming increasingly important:
  - Metadata advocates need to prove the practical value of metadata frameworks and 'outline specifications'
  - We need to move from the conceptual to the practical, need to move beyond proof-of-concept
  - Positive signs:
    - METS/NISO Z39.87
    - PREMIS WG

Issues - sustainability

- 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 ('core metadata'?)

Issues - creation and capture

- Metadata creation/capture:
  - Human agency vs. automatic capture
  - How much metadata already exists?
  - The need for automatic (or semi-automatic) capture or conversion of metadata
  - Need for metadata to be captured at creation, ingest, migration, and at other appropriate points in object life-cycle

Issues - interoperability (1)

- Interoperability is important:
  - To support the reuse of existing metadata
  - To support the exchange of digital objects between repositories
- Problems:
  - The need to cope with a wide (and growing) range of metadata standards, object types, formats, etc.
  - Growing number of repositories

Issues - interoperability (2)

- Metadata registries?
  - Provide support for the ingest process
  - May also provide support for the access function
    - The export of objects to users
    - The exchange of objects with other repositories; conversion to exchange standards
  - Help manage schema evolution
  - Possible relationship with format registries, e.g., the proposed Global File Format Registry

Summing up

- Metadata is perceived to be useful (or essential) for the long-term management of digital objects
- There is some consensus on what metadata might be required (e.g., OAIS model, specific requirements for recordkeeping, etc.)
- Less agreement on how this should be properly implemented, but there has been progress through initiatives like PREMIS and METS

Key links:

- OAIS Reference Model: http://www.ccsds.org/documents/650x0b1.pdf
- PREMIS WG: http://www.oclc.org/research/projects/pmwg/
- ERPANET Training Seminar on "Metadata in Digital Preservation" (Marburg, 2003): http://www.erpanet.org/
- Digital Curation Centre: http://www.dcc.ac.uk/
- Digital Preservation Coalition: http://www.dpconline.org/

Acknowledgements

UKOLN is funded by 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.

Also thanks to the Digital Preservation Coalition, the Digital Curation Centre, the DELOS Network of Excellence preservation cluster.