Preservation metadata

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

Session overview

- The need for preservation metadata
- Some definitions and roles
- Some influential standards
  - The OAIS Information Model
  - The PREMIS Data Dictionary
- Some final challenges

UKOLN

The need for preservation metadata

Digital preservation options (1)

- Option 1: Retaining the original media
  - Sometimes known as "technology preservation" (also keeping all necessary hardware and software)
  - However, media will decay and become obsolete (as will associated hardware and software)
  - There may be some ways to rescue content, e.g. digital archaeology or forensics (expensive and unproven)

Digital preservation options (2)

- Option 2: Retaining and maintaining the (original) bit stream
  - Known as bit-level preservation
    - An essential part of any long-term digital preservation strategy
  - However, keeping bits safe is insufficient by itself
  - There remains a need for additional information (e.g. software, hardware emulators, documentation, descriptive and contextual metadata) that supports both the authenticity of objects and their continued rendering

UKOLN

Digital preservation options (3)

- Option 3: The periodic (and ongoing) transformation of bit streams
  - Associated with migration strategies
  - Bit stream can always (?) be rendered within the current hardware and software environment
  - There still remains a need for descriptive and contextual metadata, also for additional information on the change-history of the object itself (provenance and stewardship)

Argument

- Metadata is required to support these preservation strategies:
  - All digital preservation approaches 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)
  - An important area of ongoing research and development (and increasingly implementation)

If we start with an object ...

... we will need to answer some questions about it ...

Who created this object?

What intellectual property rights are vested in the object?

What format is it?

What software will I need to render it?

Is there a codebook ?

Is this software currently available?

Are there any other dependencies?

Has the object been changed in any way since ingest?

Is the object related to other objects located within this (or any other) repository?

Who made these changes?

Can I be sure that the object is what it claims to be?

Who has previously had custody of this object?

Who currently has custody of this object?

UKOLN

Some definitions and roles

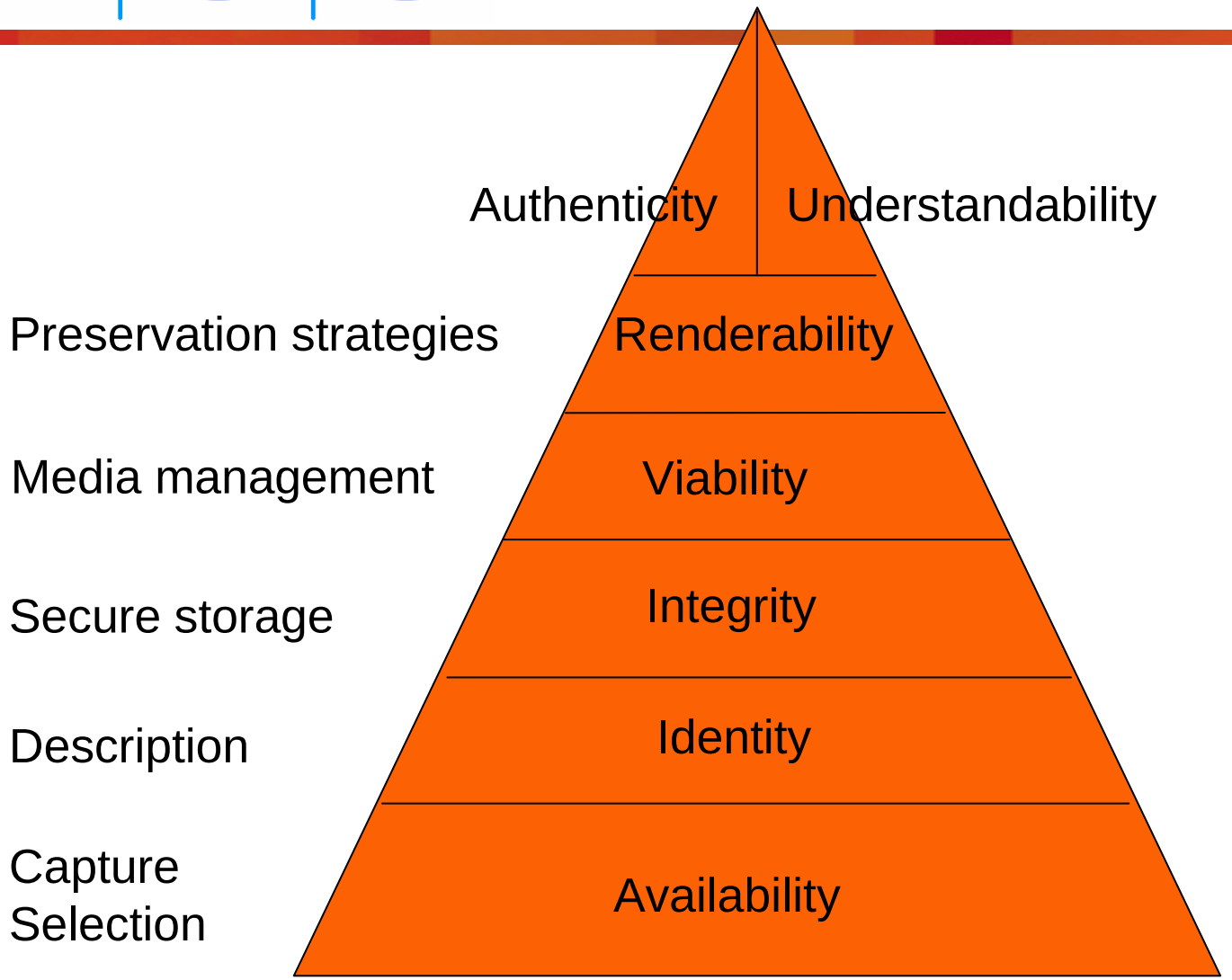
Roles of preservation metadata

- The information needed "… to find, manage, control, understand or preserve … information over time" (Adrian Cunningham, 2000)
- The various types data that will allow the re-creation and interpretation of the structure and content of digital data over time (Ludäsher, Marciano & Moore, ACM SIGMOD Record, 2001)
- The "information a repository uses to support the digital preservation process," specifically: "the functions of maintaining viability, renderability, understandability, authenticity, and identity in a preservation context" (PREMIS Data Dictionary, 2005)

UKOLN

PREMIS roles

- No specific definitions in PREMIS Data Dictionary:
  - Viability - bit streams (and systems) should be managed in a way that ensures that they continue to be available over time
  - Renderability - preservation strategies should be adopted in order to ensure that objects can be rendered in appropriate ways, e.g. within the current computing environment or through emulation
  - Understandability - Objects should be understandable (at various different levels, e.g. structure and semantics)
  - Authenticity - Objects should be what they claim to be ... bit integrity is not enough
  - Identity - Objects should be identifiable and able to be discovered in appropriate ways



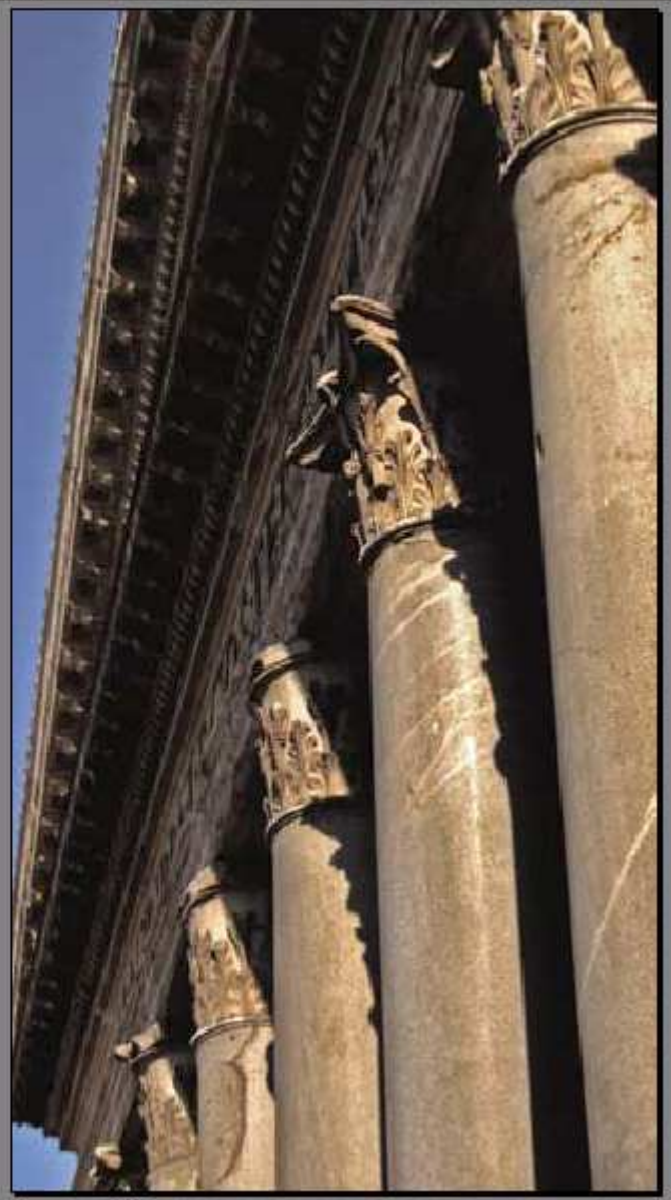
Priscilla Caplan's revised Preservation Pyramid (2005)

UKOLN

Sources of metadata (1)

- Embedded within objects themselves
  - Typical examples include TIFF headers, file properties in Office programs
  - Tools have been developed to capture some of this metadata automatically, e.g.:
    - New Zealand National Library preservation metadata extraction tool
    - JHOVE (JSTOR/Harvard Object Validation Environment) for the identification and validation of formats
  - However, can we always trust embedded metadata?
    - Do we regularly update file properties?
    - What do we do if there are conflicts?

UKOLN



Mallorca

DIRECTIONS

WRITTEN AND RESEARCHED BY

Phil Lee

ROUGH GUIDES

NEW YORK • LONDON • DELHI

www.roughguides.com

Acrobat Reader - [Rome Directions]

File Edit Document Tools View Window Help

109%

Bookmarks Thumbnails

Document Summary

File: C:\Documents and Settings\lismd\Personal\...\rome.pdf

Title: Rome Directions

Subject: Travel

Author: Martin Dunford

Keywords: Rome, Travel, Rough Guides, Directions

Binding: Left Edge

Creator: Adobe InDesign CS (3.0)

Producer: Adobe PDF Library 6.0

Created: 9/17/2004 3:04:18 PM

Modified: 1/11/2005 3:58:28 PM

File Size: 10.79 MB (11,315,377 Bytes)

Security: 40-bit RC4 (Acrobat 3.x, 4.x)

PDF Version: 1.4 (Acrobat 5.x)

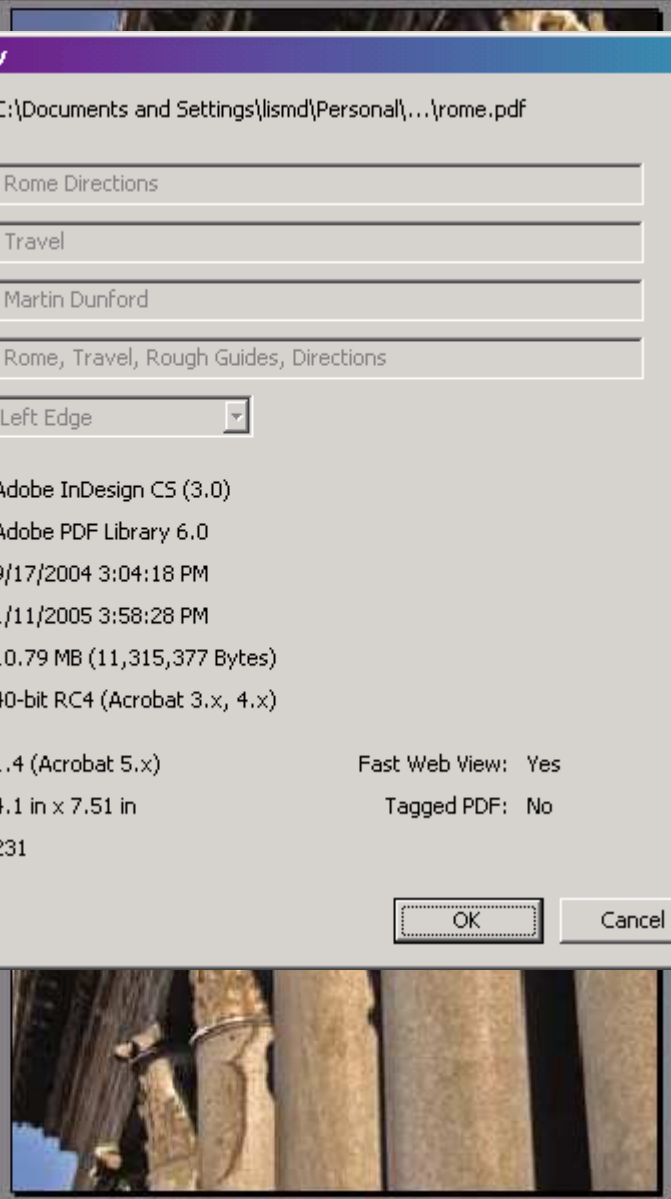
Page Size: 4.1 in x 7.51 in

Number of Pages: 231

Fast Web View: Yes

Tagged PDF: No

OK Cancel



Mallorca

DIRECTIONS

WRITTEN AND RESEARCHED BY

Phil Lee

ROUGH GUIDES

NEW YORK • LONDON • DELHI

www.roughguides.com

ii (2 of 231) 4.1 x 7.51 in

Sources of metadata (2)

- Associated with objects, e.g.:
  - Readme files or documentation
  - Databases (e.g., bibliographic catalogues, e-journal systems)
  - Documentation standards or codebooks (e.g. XML)
- Created by the preservation repository itself
  - Part of ingest process
  - Automatically captured from the ongoing management of objects, recording, e.g.:
    - Custodial history
    - Format transformations
    - Usage

UKOLN

Some influential standards

The OAIS Model

- Reference Model for an Open Archival Information System (OAIS)
  - Development managed by the Consultative Committee on Space Data Systems (CCSDS)
  - CCSDS Blue Book 650.0-B-1 (2002)
  - ISO 14721:2003 (currently under review)
  - Has established a common framework of terms and concepts
  - The information model has been very relevant to the design of preservation metadata schemas
  - Question of OAIS "conformance"

UKOLN

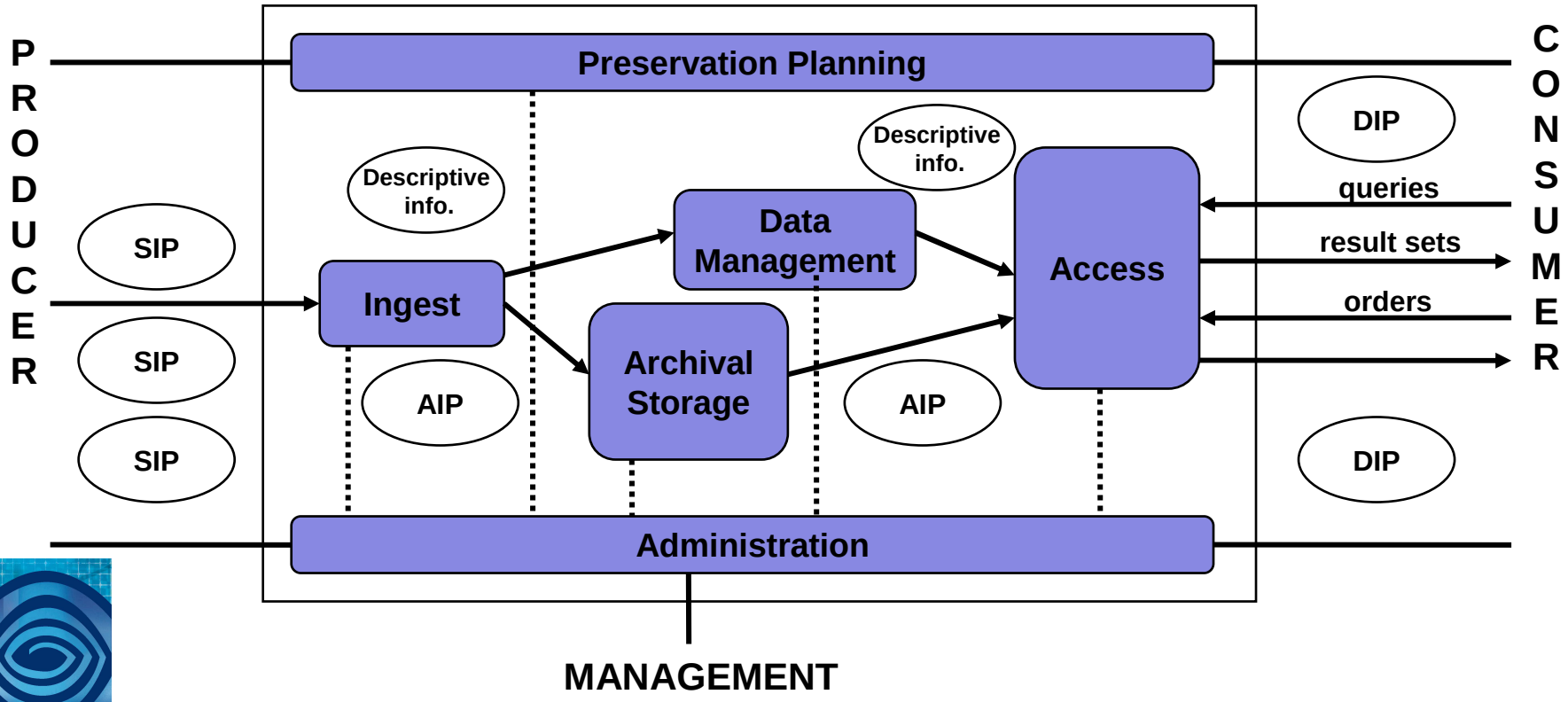
OAIS mandatory responsibilities

- Negotiating and accepting information
- Obtaining sufficient control of the information to ensure long-term preservation
- Determining the "designated community"
- Ensuring that information is **independently understandable**, i.e. without the assistance of those who produced it
- Following documented policies and procedures
- Making the preserved information available

OAIS functional model (1)

- Six entities
  - Ingest
  - Archival Storage
  - Data Management
  - Administration
  - Preservation Planning
  - Access
- Described using UML diagrams ...

OAIS functional model (2)



OAIS Functional Entities (Figure 4-1)

OAIS information objects

- Information Object (basic concept)
  - Data Object (bit-stream)
  - Representation Information
    - Permits “the full interpretation of Data Object into meaningful information”
    - Includes documentation, software, metadata, etc.
- Information Object Classes
  - Content Information
  - Preservation Description Information (PDI)
  - Packaging Information
  - Descriptive Information

UKOLN

OAIS information packages

- Information package:
  - Container that encapsulates Content Information and Preservation Description Information (PDI)
  - Different packages defined for submission to an archive (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"
  - PDI = other information "which will allow the understanding of the Content Information over an indefinite period of time"
    - Reference, Provenance, Context, Fixity

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

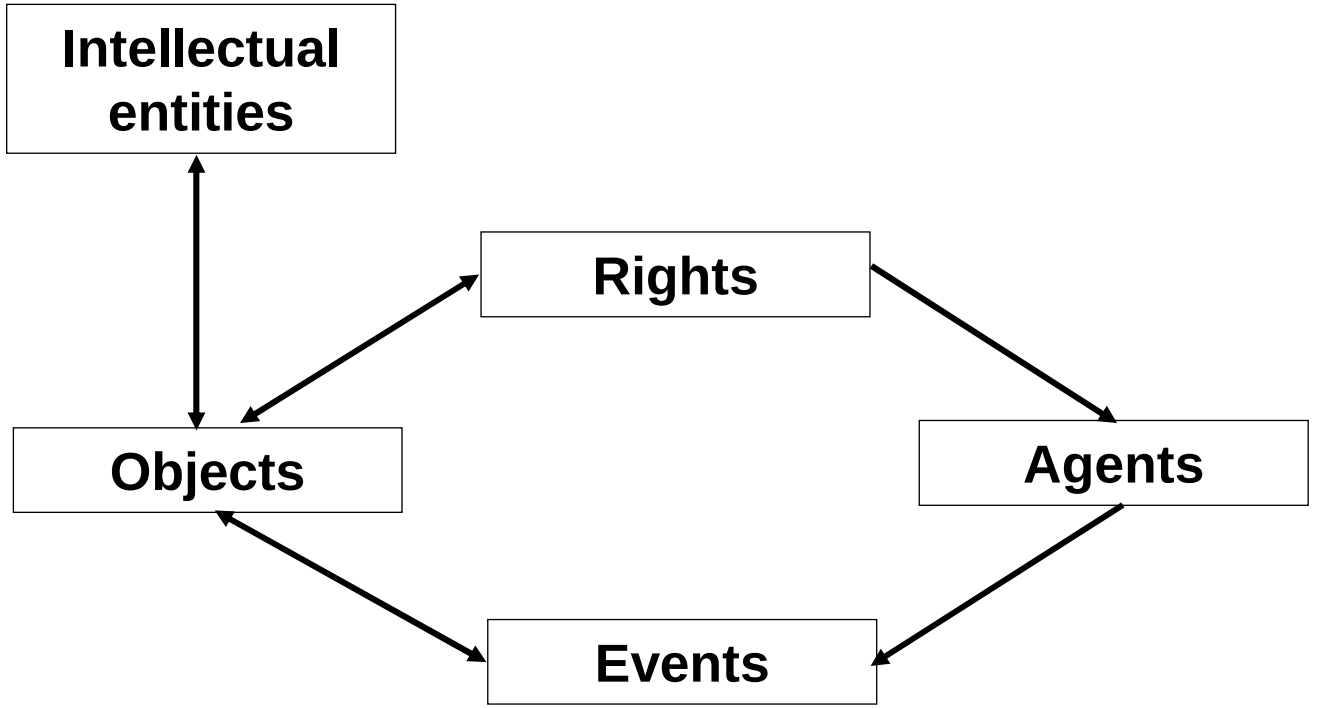
UKOLN

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 (September 2004)
  - PREMIS Data Dictionary (May 2005)
    - The data dictionary is a translation of the OAIS-based 2002 *Framework* into a set of implementable semantic units
    - Based on data model ...

UKOLN

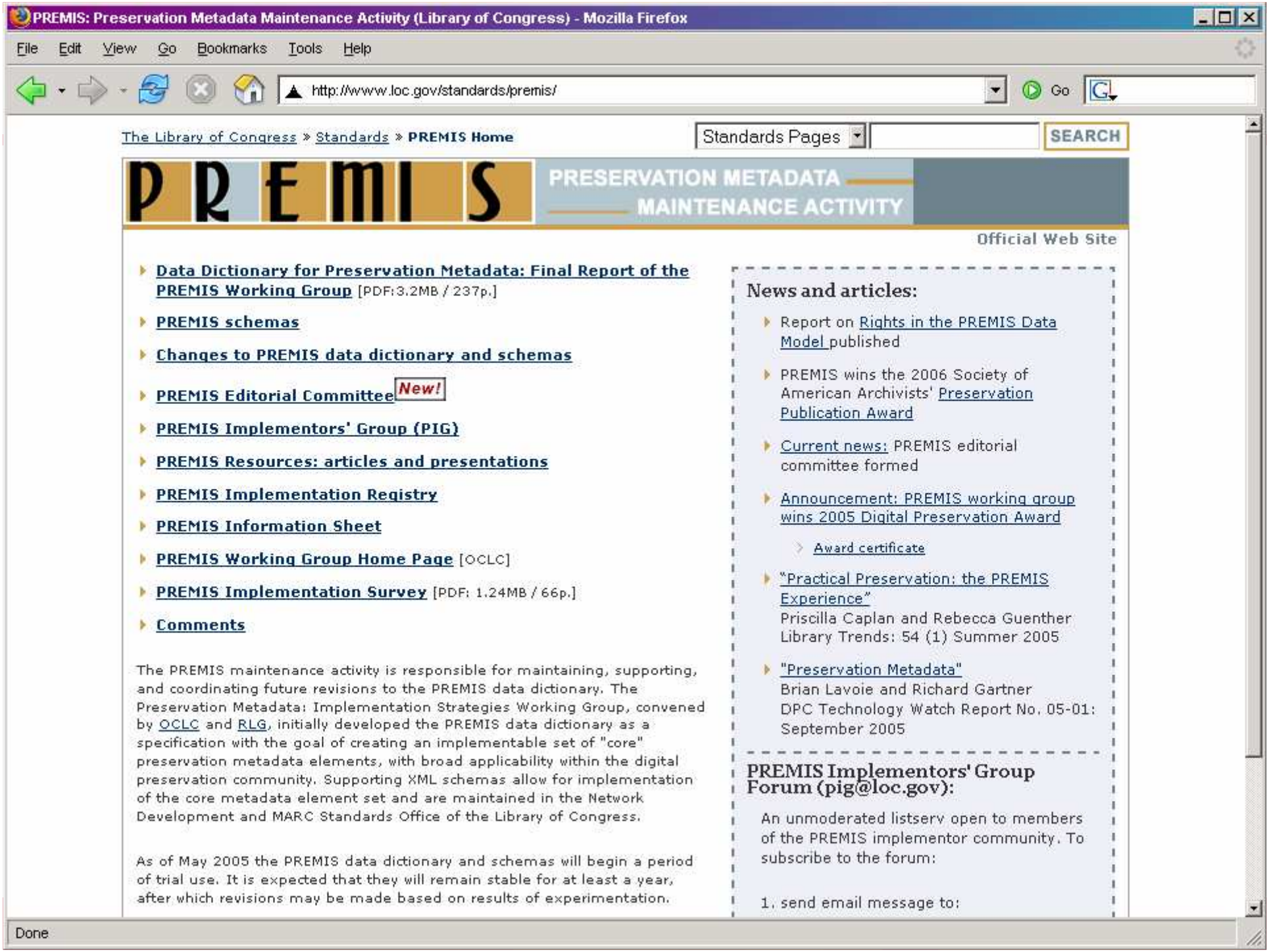
PREMIS data model



UKOLN

PREMIS Data Dictionary

- Defined various "semantic units" for:
  - Objects ... at three levels of entity (representation, file, bitstream)
  - Events ... metadata about actions
  - Agents ... but are not the main focus
  - Rights ... primarily those that relate to preservation
- Also:
  - An XML implementation
- Maintenance activity (led by the Library of Congress)
- PREMIS Implementors' Group (PIG)
- Already thinking about lessons for version 2.0


PREMIS: Preservation Metadata Maintenance Activity (Library of Congress) - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.loc.gov/standards/premis/
Go
The Library of Congress » Standards » PREMIS Home
Standards Pages
SEARCH
PREMIS
PRESERVATION METADATA MAINTENANCE ACTIVITY
Official Web Site
Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group [PDF:3.2MB / 237p.]
PREMIS schemas
Changes to PREMIS data dictionary and schemas
PREMIS Editorial Committee New!
PREMIS Implementors' Group (PIG)
PREMIS Resources: articles and presentations
PREMIS Implementation Registry
PREMIS Information Sheet
PREMIS Working Group Home Page [OCLC]
PREMIS Implementation Survey [PDF: 1.24MB / 66p.]
Comments
The PREMIS maintenance activity is responsible for maintaining, supporting, and coordinating future revisions to the PREMIS data dictionary. The Preservation Metadata: Implementation Strategies Working Group, convened by OCLC and RLG, initially developed the PREMIS data dictionary as a specification with the goal of creating an implementable set of "core" preservation metadata elements, with broad applicability within the digital preservation community. Supporting XML schemas allow for implementation of the core metadata element set and are maintained in the Network Development and MARC Standards Office of the Library of Congress.
As of May 2005 the PREMIS data dictionary and schemas will begin a period of trial use. It is expected that they will remain stable for at least a year, after which revisions may be made based on results of experimentation.
News and articles:
Report on Rights in the PREMIS Data Model published
PREMIS wins the 2006 Society of American Archivists' Preservation Publication Award
Current news: PREMIS editorial committee formed
Announcement: PREMIS working group wins 2005 Digital Preservation Award
Award certificate
"Practical Preservation: the PREMIS Experience"
Priscilla Caplan and Rebecca Guenther
Library Trends: 54 (1) Summer 2005
"Preservation Metadata"
Brian Lavoie and Richard Gartner
DPC Technology Watch Report No. 05-01: September 2005
PREMIS Implementors' Group Forum (pig@loc.gov):
An unmoderated listserv open to members of the PREMIS implementor community. To subscribe to the forum:
1. send email message to:
Done

Other standards

- Standards developed from many different perspectives:
  - PREMIS Data Dictionary
  - OCLC/RLG Preservation Metadata Framework, Cedars, NEDLIB, NLA, NLNZ ... OAIS influence has been strongest in this area
  - METS, NISO Z39.87 (emerged from digitisation contexts)
- Other standards have also been developed with other aspects of object management in mind:
  - Records management (VERS, RKMS, ISO 23081-1 Metadata for Records)
  - Multimedia (MPEG-7, SMPTE Metadata Dictionary)
  - Rights management (MPEG-21)

Some final challenges

Is metadata sustainable?

- Metadata is expensive to create and maintain:
  - There is a need to balance the risks of data loss (or costs of recovery) with the costs of creating metadata
  - Automatic capture of some types of metadata
    - Metadata already embedded in objects or in secondary databases; also need to capture event metadata from archive processes
  - Sharing information via registries of format information (Representation Information)
  - Avoid imposing unnecessary costs:
    - Need to identify the *right* metadata (or 'core metadata')

The role of shared registries

- For the sharing of information about formats (and metadata)
  - There is "… a pressing need to establish reliable, sustained repositories of file format specifications, documentation, and related software" (Lawrence, *et al*., 2000)
  - Examples:
    - Global Digital Format Registry (GDFR)
      - Harvard University Library
      - Funded by the Andrew W. Mellon Foundation
    - PRONOM technical registry (The National Archives)
    - DCC Representation Information Registry and Repository (demo)

UKOLN

Representation Information Registry Repository - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://registry.dcc.ac.uk/omar/

Go

Representation Information Registry Repository

Current User: Registry Guest

Home Frequently Asked Questions Documentation User Guide

Login Reset Locale End Session Versioning ON

Content Language: English (United States)

Tasks Search Explore

Create User Account

Create a New Registry Object

Customize

Welcome to the Representation Information Web UI Registry Browser

This registry repository curates OAIS reference model (ISO:14721:2002) defined Representation Information which is intended to add meaning to data and aid its long-term preservation.

Objectives

1. to create a centralized site for the sharing of OAIS defined Representation Information
2. to promote the use of Representation Information in digital curation and long-term preservation of data.

Getting Started

To start browsing Registry content, click on either the Search or Explore links in the left sidebar. The Search link allows you to search for content using form-based queries. The Explore link gives you a filesystem view of Registry content.

Done

Further reading

- Three chapters from the DCC Digital Curation Manual: http://www.dcc.ac.uk/resource/curation-manual/chapters/
  - Priscilla Caplan, "Preservation metadata" (July 2006)
  - Wendy Duff & Marlene van Ballegooie, "Archival Metadata" (May 2006)
  - Michael Day, "Metadata" (November 2005)
- Brian Lavoie & Richard Gartner, "Preservation metadata." DPC Technology Watch Report 05-01 (September 2005): http://www.dpconline.org/docs/reports/dpctw05-01.pdf
- PREMIS Data Dictionary for Preservation Metadata (May 2005): http://www.oclc.org/research/projects/pmwg/
- Reference Model for an Open Archival Information System (OAIS), CCSDS 650 0-B-1 (January 2002): http://public.ccsds.org/publications/archive/650x0b1.pdf

UKOLN

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 e-Science Programme: http://www.dcc.ac.uk/