Seminar: OAIS Model application in digital preservation projects

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

Preservation of Digital Heritage: Basic Concepts and Main Initiatives, Madrid, 14-16 March 2006

Seminar outline

- Introduction to the OAIS Model:
  - Background
  - Mandatory Responsibilities
  - Functional Model
  - Information Model
- Main application areas:
  - Repository compliance
  - The analysis and comparison of repositories
  - Informing system design
  - Preservation metadata

OAIS background

- Reference Model for an Open Archival Information System (OAIS)
  - Nothing to do with the OAI (Open Archives Initiative) or OAI-PMH
  - 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
  - http://public.ccsds.org/publications/archive/650x0b1.pdf

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

Preservation of Digital Heritage, Madrid, 14-16 March 2006

OAIS definitions (1)

- Provides definitions of terms, e.g.:
  - **OAIS** - "An archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community"
  - **Designated Community** - the community of stakeholders and users that the OAIS serves
  - **Knowledge Base** - a set of information, incorporated by a user or system, that allows that user or system to understand the received information

OAIS definitions (2)

- **Information Object** - Data Object + Representation Information
- **Representation Information** - any information required to render, interpret and understand digital data
- **Information Package** - Conceptual linking of Content Information + Preservation Description Information + Packaging Information (Submission, Archival and Dissemination Information Packages)
- **Preservation Description Information** - information (metadata) about Provenance, Context, Reference, Fixity information

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

Preservation of Digital Heritage, Madrid, 14-16 March 2006

OAIS high level concepts (1)

- – The *environment* of an OAIS (Producers, Consumers, Management)
- – Definitions of *information*, Information Objects and their relationship with Data Objects
- – Definitions of *Information Packages*, conceptual containers of Content Information and Preservation Description Information
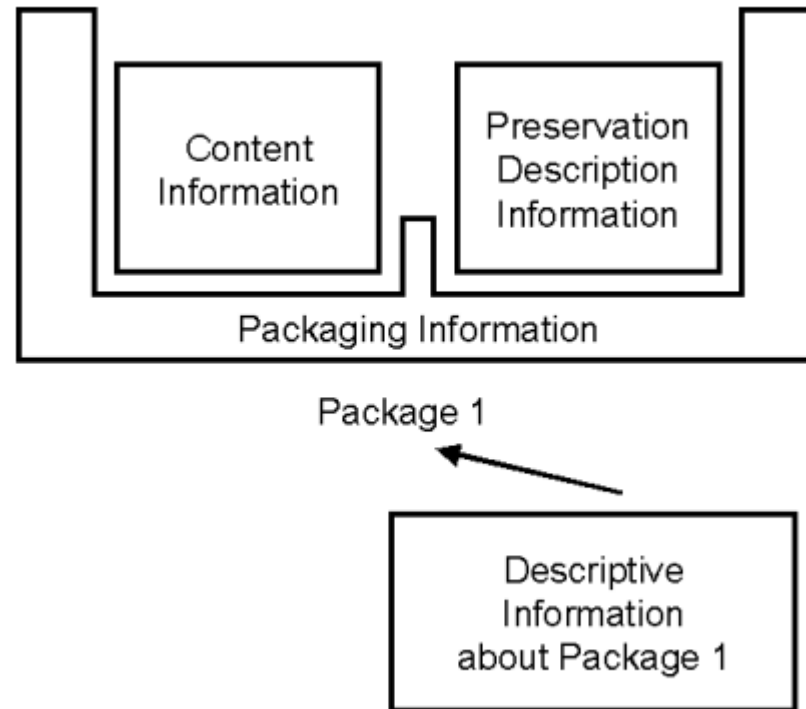
OAIS high level concepts (2)

Information Package Concepts and Relationships (Figure 2-3)

OAIS mandatory responsibilities (1)

- Negotiate for and accept appropriate information from information Producers
- Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation
- Determine, either by itself or in conjunction with other parties, which communities should become the *Designated Community* and, therefore, should be able to understand the information provided

OAIS mandatory responsibilities (2)

- Ensure that the information to be preserved is *Independently Understandable* to the Designated Community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information

- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original

- Make the preserved information available to the Designated Community

OAIS Functional Model (1)

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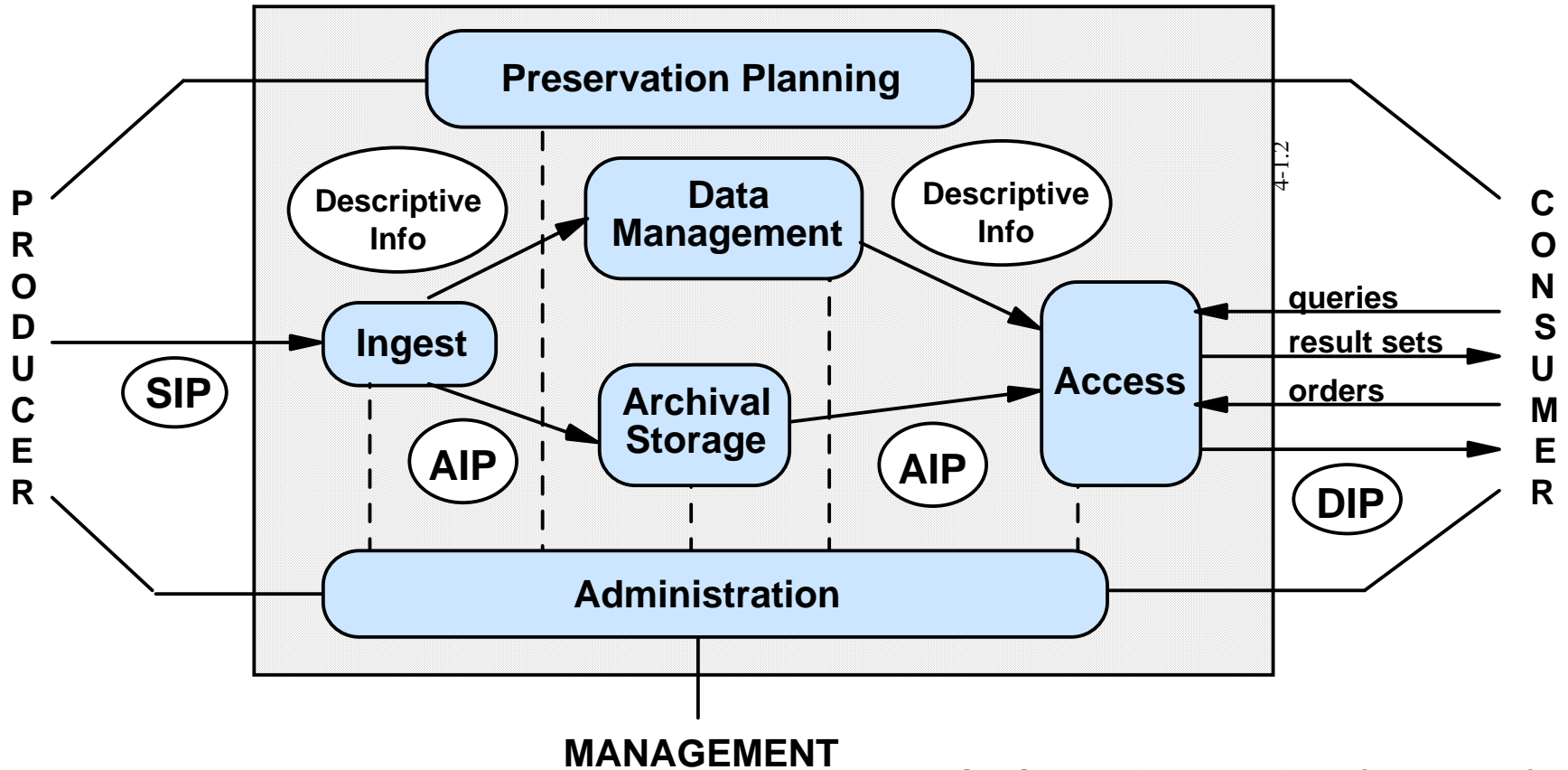 - Access
- Described using UML diagrams ...

OAIS Functional Model (2)

Preservation Planning

Descriptive Info | **Data Management** | **Descriptive Info**

PRODUCER

Ingest

SIP

AIP

Archival Storage

AIP

Access

queries
result sets
orders

CONSUMER

DIP

Administration

MANAGEMENT

OAIS Functional Entities (Figure 4-1)

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

Preservation of Digital Heritage, Madrid, 14-16 March 2006

OAIS Functional Entities (1)

- **Ingest** - services and functions that accept SIPs from Producers; prepares AIPs for storage, and ensures that AIPs and their supporting Descriptive Information become established within the OAIS

- **Archival Storage** - services and functions used for the storage and retrieval of AIPs

Functions of Archival Storage

OAIS Functional Entities (2)

- **Data Management** -services and functions for populating, maintaining, and accessing a wide variety of information

- **Administration** - services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis

- **Preservation Planning** - services and functions for monitoring the OAIS environment and ensuring that content remains accessible to the Designated Community
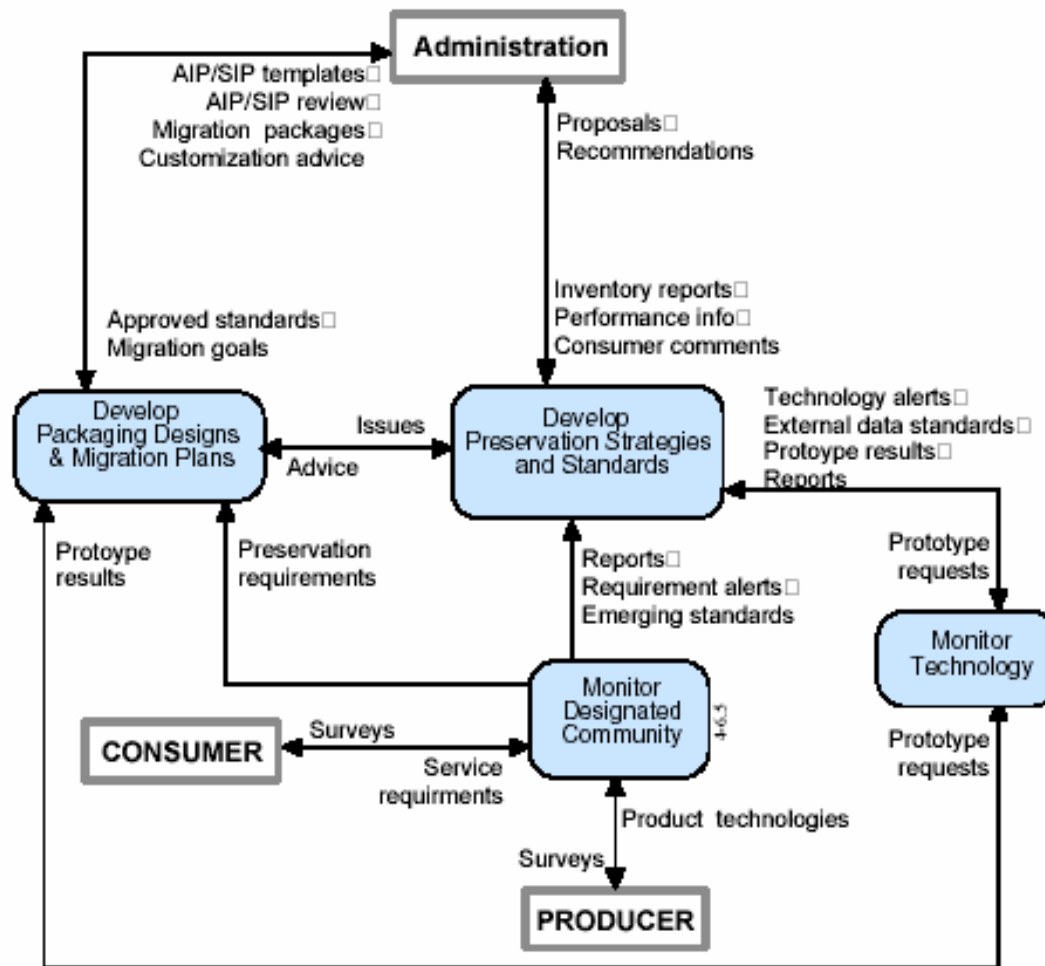
Preservation Planning Functions

Preservation of Digital Heritage, Madrid, 14-16 March 2006

OAIS Functional Entities (3)

- **Access** - services and functions which make the archival information holdings and related services visible to Consumers

OAIS Information Objects (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
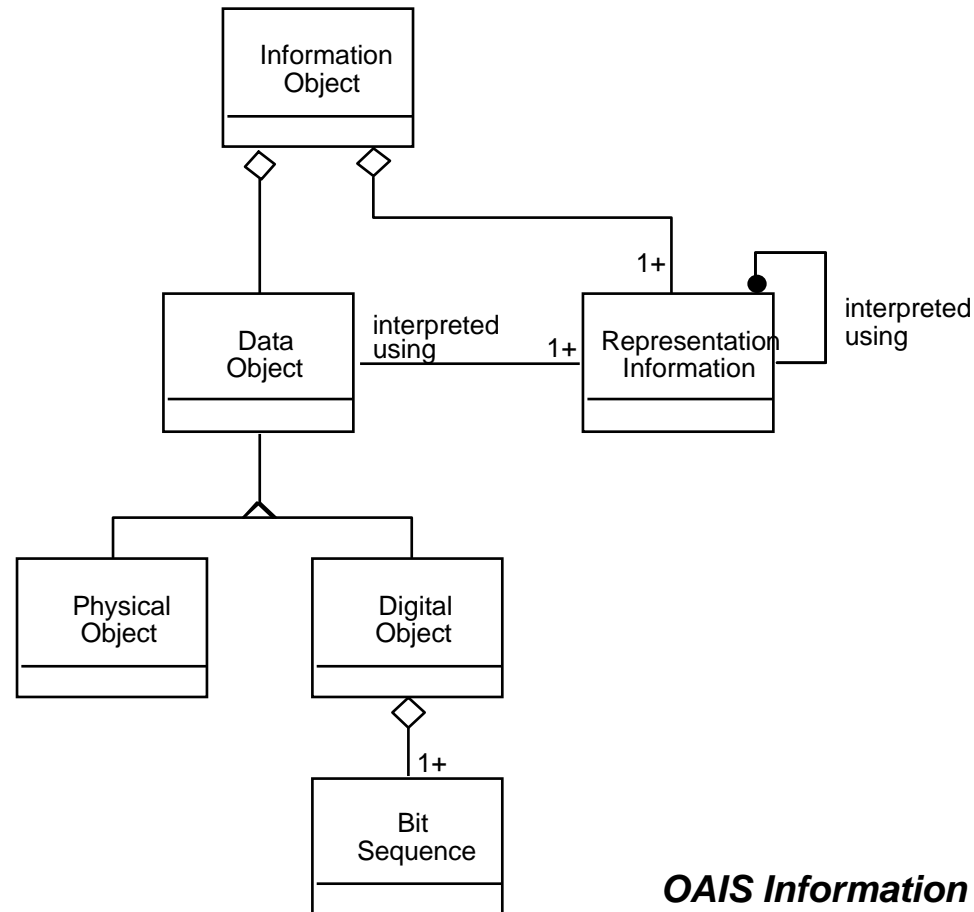- Descriptive Information

OAIS Information Objects (2)

OAIS Information Object (Figure 4-10)

OAIS Information Objects (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
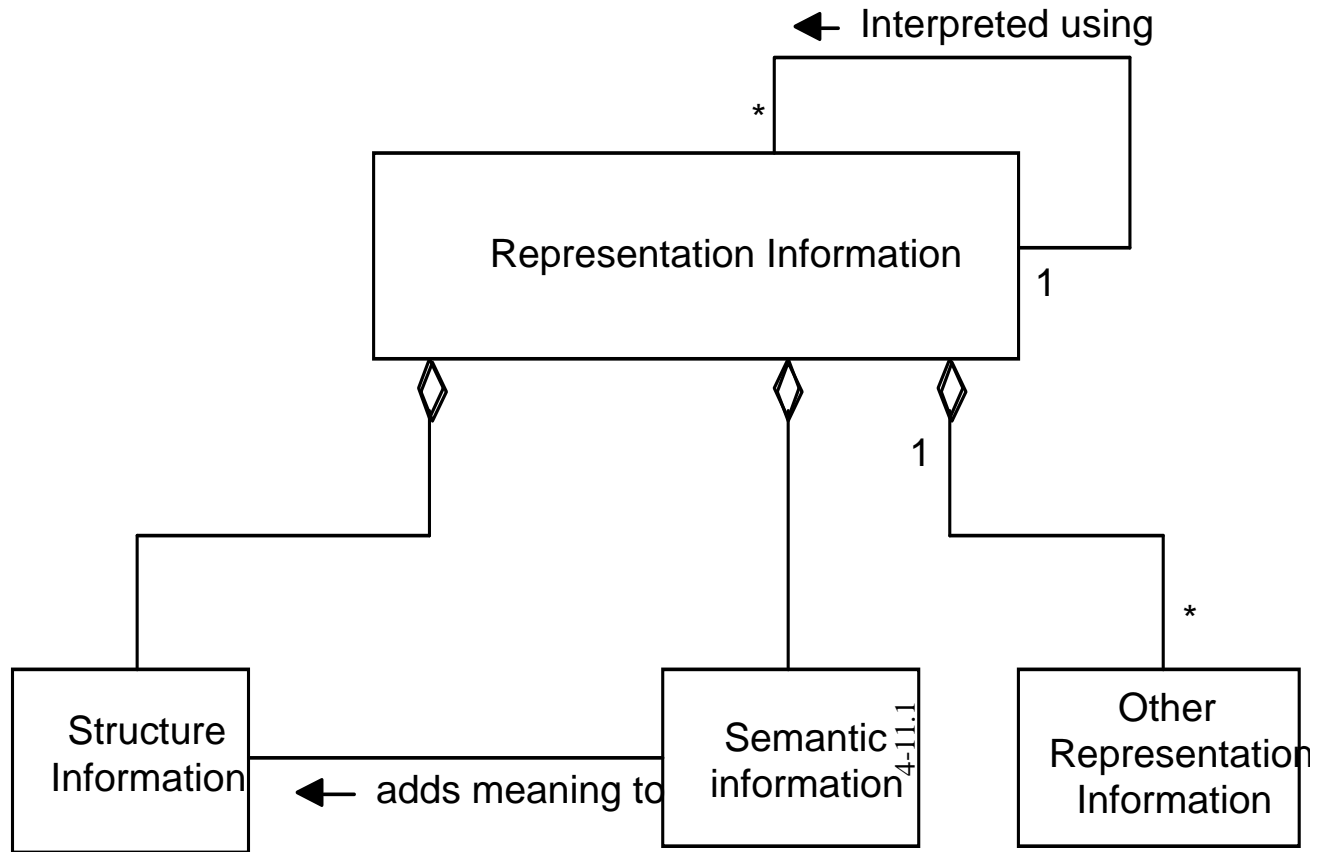   - e.g. Format registries (GDFR, PRONOM)

OAIS Information Objects (4)

OAIS Representation Information Object (Figure 4-11)

OAIS Information Packages (1)

- 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 Packages (2)

- 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)
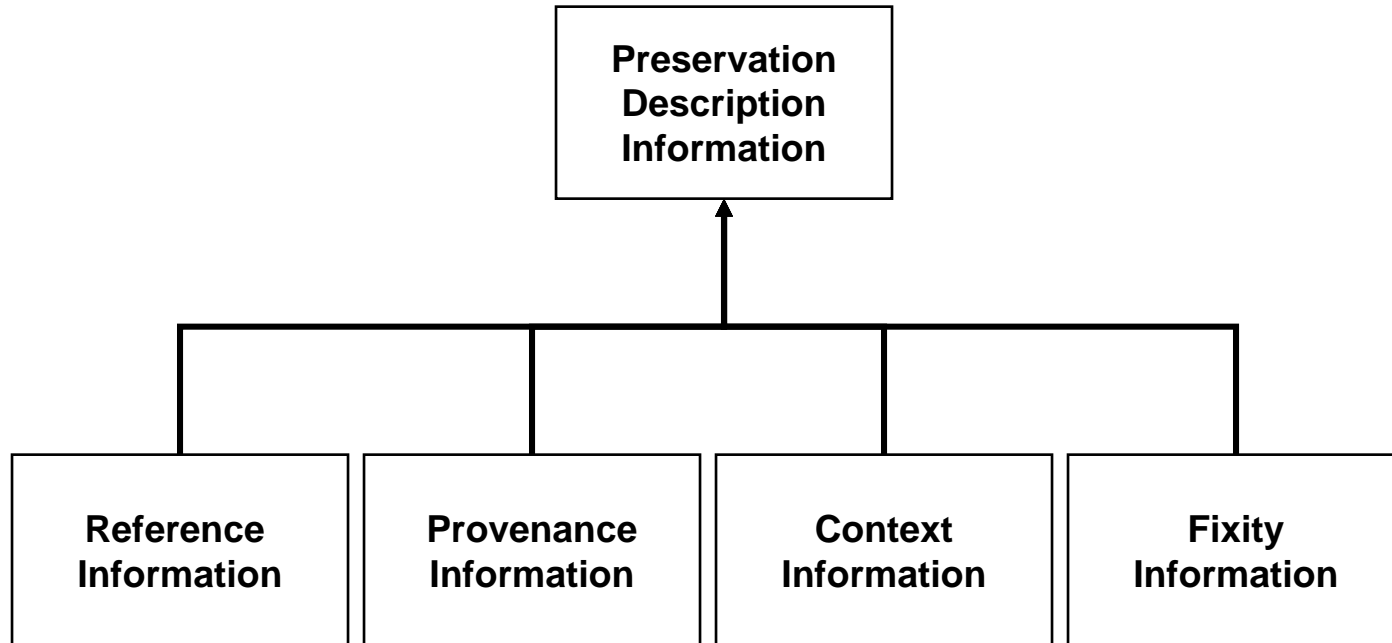
OAIS Information Packages (3)

PDI Preservation Description Information (Figure 4-16)

OAIS Information Packages (4)

- **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
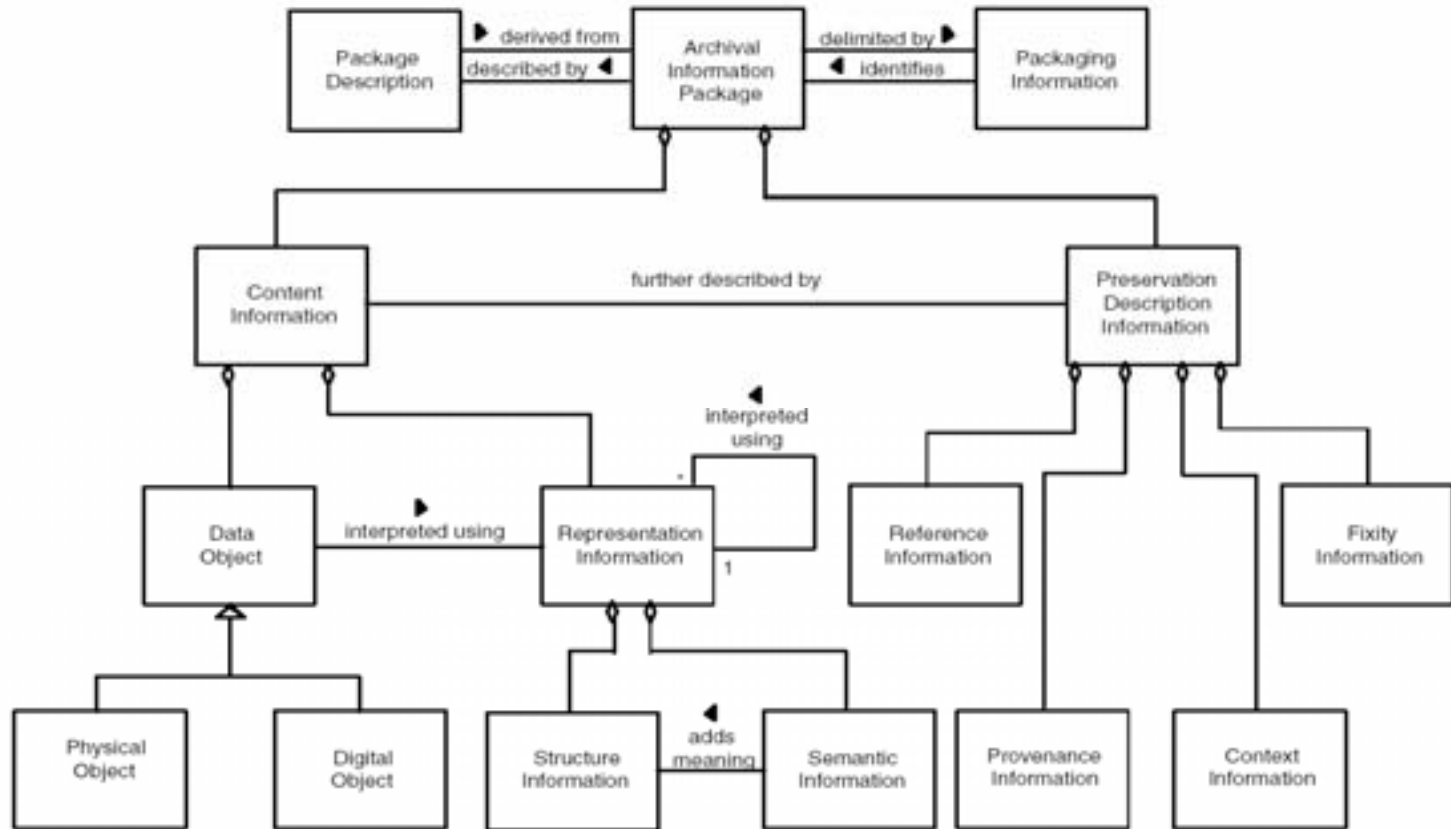
UKOLN

UNIVERSITY OF BATH

D|C|C

OAIS Information Packages (4)

OAIS Information Model

- – Also defines:
  - Archival Information Units and Archival Information Collections
    - – Recognises the complexity some some objects, addresses granularity
  - Information Package transformations
    - – For Ingest and Access

OAIS - other perspectives

– Preservation

- Migration, e.g refreshment, replication, repackaging, transformation

- Preservation of look and feel (e.g., emulation, virtual machines)

– Archive interoperability

- Interaction between OAIS archives (e.g., co-operating and federated archives)

– Examples of existing archives (annex)

Implementing the OAIS model

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

Fundamentals of implementation (1)

- OAIS is a reference model (conceptual framework), NOT a blueprint for system design
- It informs the design of system architectures, the development of systems and components
- It provides common definitions of terms … a common language, means of making comparison
- But it does NOT ensure consistency or interoperability between implementations

Fundamentals of implementation (2)

- ISO 14721:2003
  - Follows the Recommendation made available by the CCSDS
  - However, earlier versions of the model made available by the CCSDS informed implementations long before its issue by ISO
- Main areas of influence:
  - Compliance and certification
  - Analysis and comparison of archives
  - Informing system design
  - Preservation metadata

Conformance and certification

OAIS conformance (1)

- Many repositories or preservation tools claim OAIS influence or compliance:
  - e.g., DSpace, OCLC Digital Archive, METS
  - LOCKSS System has produced a "formal statement of conformance to ISO 14721:2003" (lockss.stanford.edu/)
- The OAIS model claims to be a basis for conformance (OAIS 1.4), e.g.:
  - Supporting the information model (OAIS 2.2),
  - Fulfilling mandatory responsibilities (OAIS 3.1)

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

Preservation of Digital Heritage, Madrid, 14-16 March 2006

OAIS conformance (2)

- 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*
  - Following documented policies and procedures
  - Making the preserved information available

Trusted digital repositories (1)

- OCLC/RLG Digital Archive Attributes Working Group
  - Trusted Digital Repositories report (2002)
    - http://www.rlg.org/legacy/longterm/repositories.pdf
  - Recommended the development of a process for the certification of digital repositories
    - Audit model
    - Standards model
  - Goes well beyond OAIS mandatory responsibilities …

Trusted digital repositories (2)

- Identified specific attributes:
  - Compliance with OAIS
  - Administrative responsibility
  - Organisational viability
  - Financial sustainability
  - Technological and procedural suitability
  - System security
  - Procedural accountability

RLG-NARA Task Force (1)

- RLG-NARA Task Force on Digital Repository Certification
  - Supported by RLG and the US National Archives and Records Administration (NARA)
  - To define certification model and process
    - Identify those things that need to be certified (attributes, processes, functions, etc.)
    - Develop a certification process (organisational implications)
  - *An audit checklist for the certification of trusted digital repositories* (draft, August 2005)

RLG-NARA Task Force (2)

- Audit checklist criteria:
  - Organizational:
    - Governance and organizational viability, Organizational structure and staffing, Procedural accountability and policy framework, Financial sustainability, Contracts, licenses and liabilities
  - Repository functions
    - Follows OAIS Functional Model
  - Designated Community and the usability of information
  - Technologies and technical infrastructure

RLG-NARA Task Force (3)

- Checklist intended to be used both for:
  - Self evaluation
  - An independently administered audit
- Provides a framework for certification and documentation of repository practice …

RLG-NARA Task Force (4)

The screenshot shows a PDF checklist document with the following table:

	Planned?	Documented?	Implemented?	Evaluated?	Notes
A. The Organization					
A1. Governance & organizational viability					
A1.1. Repository has a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information on behalf of depositors.					
A1.2. Repository has a formal succession plan, contingency plans, and/or escrow arrangements in place in case repository ceases to operate or substantially changes its scope (i.e., return with adequate prior notification of digital objects to depositors and/or trusted inheritors identified).					
A2. Organizational structure & staffing					
A2.1. Repository staff have skills and expertise appropriate to their duties.					
A2.2. Repository has appropriate number of staff to support all functions and services designated in agreements with depositors.					
A2.3. Repository commits to professional development to keep staff expertise and skills current.					
A3. Procedural accountability & policy framework					
A3.1. Repository has a mechanism in place for reviewing, updating, and developing comprehensive policies and procedures as repositories grow and as the community practice evolves.					
A3.2. Repository has monitoring and feedback mechanisms in place to ensure continued operation, support problem resolution, and address evolving requirements of providers and consumers.					
A3.3. Repository is committed to formal, periodic review and assessment to ensure continued development.					
A3.4. Repository has a documented history of the changes to its operations, procedures, software, hardware, traceable to its preservation strategies where appropriate.					
A3.5. Repository commits to transparency and accountability in all actions supporting the operation and management of the repository.					

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

Preservation of Digital Heritage, Madrid, 14-16 March 2006

CRL Certification project

- Center for Research Libraries (CRL) Certification of Digital Archives project
- Funded by the Andrew W. Mellon Foundation
- Builds on RLG-NARA WG work to further develop certification processes and metrics
- Develop profile and business model for a certifying agency
- Participating archives:
  - Koninklijke Bibliotheek, Portico, Inter-university Consortium for Political and Social Research, LOCKSS, …

The analysis and comparison of repositories

The analysis of existing services

– A process started in the annexes to the model itself

– Looking at existing services and processes, mapping them to OAIS functional and information model

– Main uses:

  • Identifying significant gaps

  • Provides a common language for the comparison of archives

BADC/APS case study (1)

- British Atmospheric Data Centre
  - A data centre of the Natural Environment Research Council (NERC)
  - Evaluating the use of the CCLRC's Atlas Petabyte Storage (APS) Service for long-term data storage
  - Mapping OAIS to combined BADC/APS
    - BADC responsible for Ingest and Access
    - APS responsible for Archival Storage
    - Jointly responsible for Data Management and Administration

BADC/APS case study (2)

- Application of OAIS revealed:
  - Feedback on how well the BADC/APS fulfilled OAIS mandatory responsibilities
  - AIP needs better definition
  - Weaknesses identified with the Preservation Planning role, e.g. little explicit monitoring of technology or the Designated Community
- OAIS helps to identify limitations
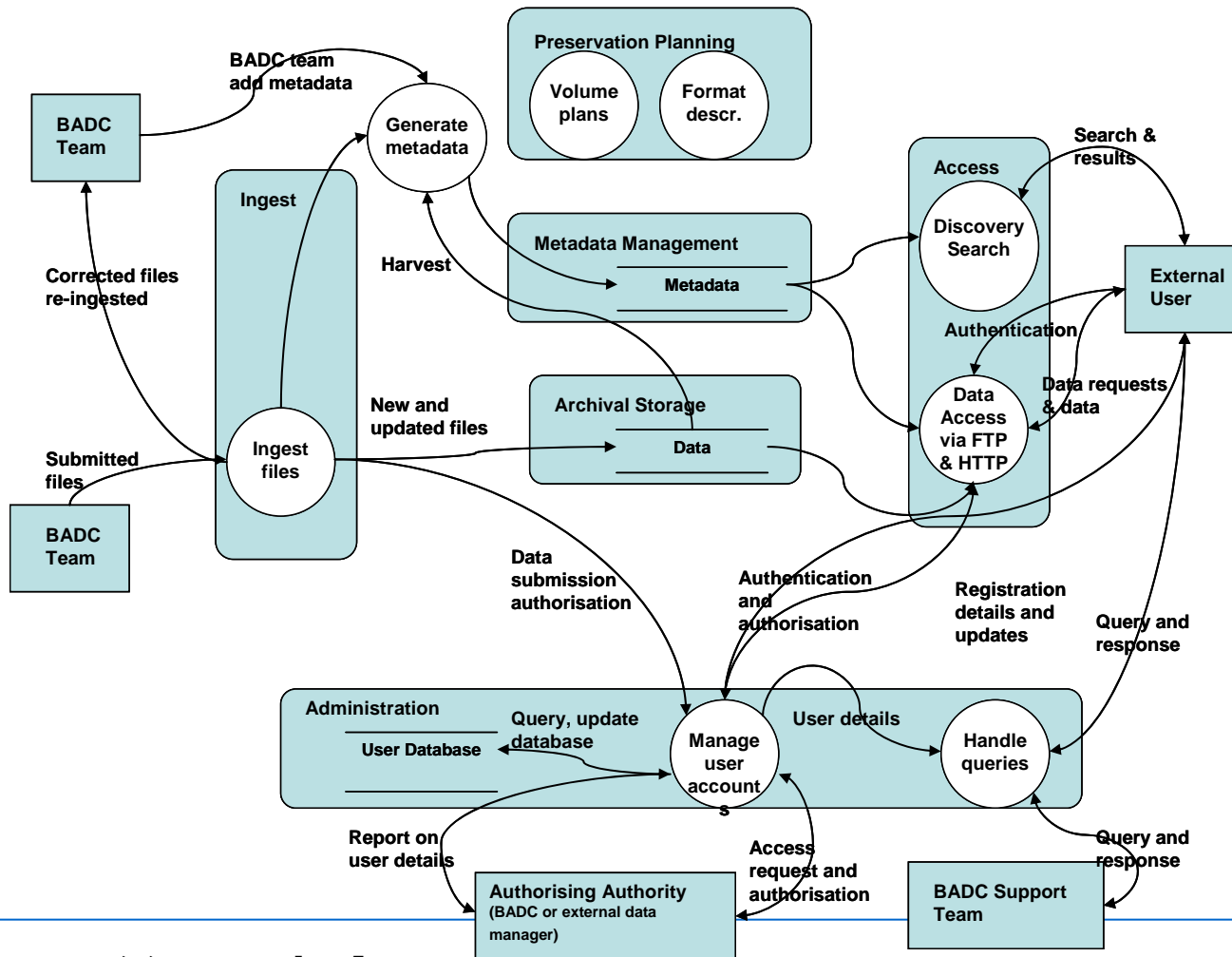- For more details, see: Corney, *et al.* (2004) http://www.allhands.org.uk/2004/proceedings/papers/156.pdf

BADC/APS case study (3)

UKDA and TNA case study (1)

- UK Data Archive and The National Archives

  - JISC-funded project mapping UKDA and TNA to OAIS functional and information models

  - Published in: Beedham, *et al*., (2005). http://www.data-archive.ac.uk/news/ publications/oaismets.pdf

UKDA and TNA case study (2)

- Conclusions:
  - Noted that there was no existing methodology for testing OAIS compliance
    - Recommended the production of guidelines or manual
  - The OAIS Mandatory Responsibilities are carried out by almost any archive
  - The OAIS Designated Community concept assumes a identifiable and relatively homogenous user community; this is not the case for either UKDA or TNA

UKDA and TNA case study (3)

- – Conclusions (continued):
  - The relationship between AIPs and DIPs needs clarification
  - The OAIS Administration function may be difficult for small archives to fulfil adequately
  - Model not scalable - report proposes an 'OAIS Lite'
  - Information categories (e.g. PDI) are too general to allow mapping of metadata elements from other schemas (p. 70)

UKDA and TNA case study (4)

– Conclusions (continued):

- But ... OAIS terminology *was* useful to support communication between UKDA and TNA

Informing system design

Informing system design (1)

- OAIS is not a blueprint for system design
    - "It is assumed that implementers will use this reference model as a guide while developing a specific implementation to provide identified services and content" (OAIS 1.4)
- But it has been used to inform the design of systems
    - This can be difficult because the model does not distinguish between management and technical processes
    - Need to first identify the areas that can be supported by technical development

Informing system design (2)

- Many examples:
  - Complete systems:
    - aDORe (Los Alamos National Laboratory)
    - OCLC Digital Archive Service
    - Stanford Digital Repository
    - MathArc (Cornell UL and SUB Göttingen)
  - Tools:
    - Dspace, FEDORA, …
    - DCC Representation Information Registry
    - Harvard University Library XML-based Submission Information Package for e-journal content

Informing system design (3)

– As a basis for domain-specific modelling

- InterPARES project Preservation Task Force
- Preserve Electronic Records model
  - Formally modelled the specific processes and functions involved with preserving electronic records
  - Developed "… a specification of an OAIS for the specific classes of information objects comprising electronic records and archival aggregates of such records"
  - http://www.interpares.org/

Preservation metadata

Preservation metadata (1)

- Metadata:
  - Data about data
  - Structured information about objects that supports various types of activity: discovery, retrieval, management, etc.
  - Often divided into descriptive, structural and administrative categories

- Preservation metadata
  - The information a repository uses to support the digital preservation process" (PREMIS WG)
  - Cuts across all metadata categories

Preservation metadata (2)

- The OAIS Information Model has been used to inform the development of many preservation metadata schemas, e.g.:
  - Draft schemas developed by the National Library of Australia, Cedars project, NEDLIB project, etc.
  - METS (Metadata Encoding and Transmission Standard) interpreted as an implementation of the OAIS Information Package concept
  - Information Model explicitly used for the structure of the OCLC/RLG Metadata Framework (2002)
  - A slightly different approach has been taken by the PREMIS Working Group

PREMIS Working Group (1)

- Working Group on Preservation Metadata: Implementation Strategies
  - Supported by OCLC and RLG
  - Established in 2003
  - International working group and advisory committee
  - Chairs: Priscilla Caplan and Rebecca Guenther

PREMIS Working Group (2)

- Building on older activity:
  - Working Group on Preservation Metadata (2000-02)
    - Preservation Metadata Framework (June 2002)
    - Explicitly based on the OAIS Information Model
- PREMIS objectives:
  - A 'core' set of preservation metadata elements (Data Dictionary)
  - Strategies for encoding, packaging, storing, managing, and exchanging metadata

PREMIS Working Group (3)

- Main PREMIS outputs:
  - Implementation Survey report (September 2004)
    - Based on ~50 responses
    - Snapshot of practice, noting trends
  - PREMIS Data Dictionary 1.0 (May 2005)
    - 237 pp.

  - All WG documents are available from: http://www.oclc.org/research/projects/pmwg/
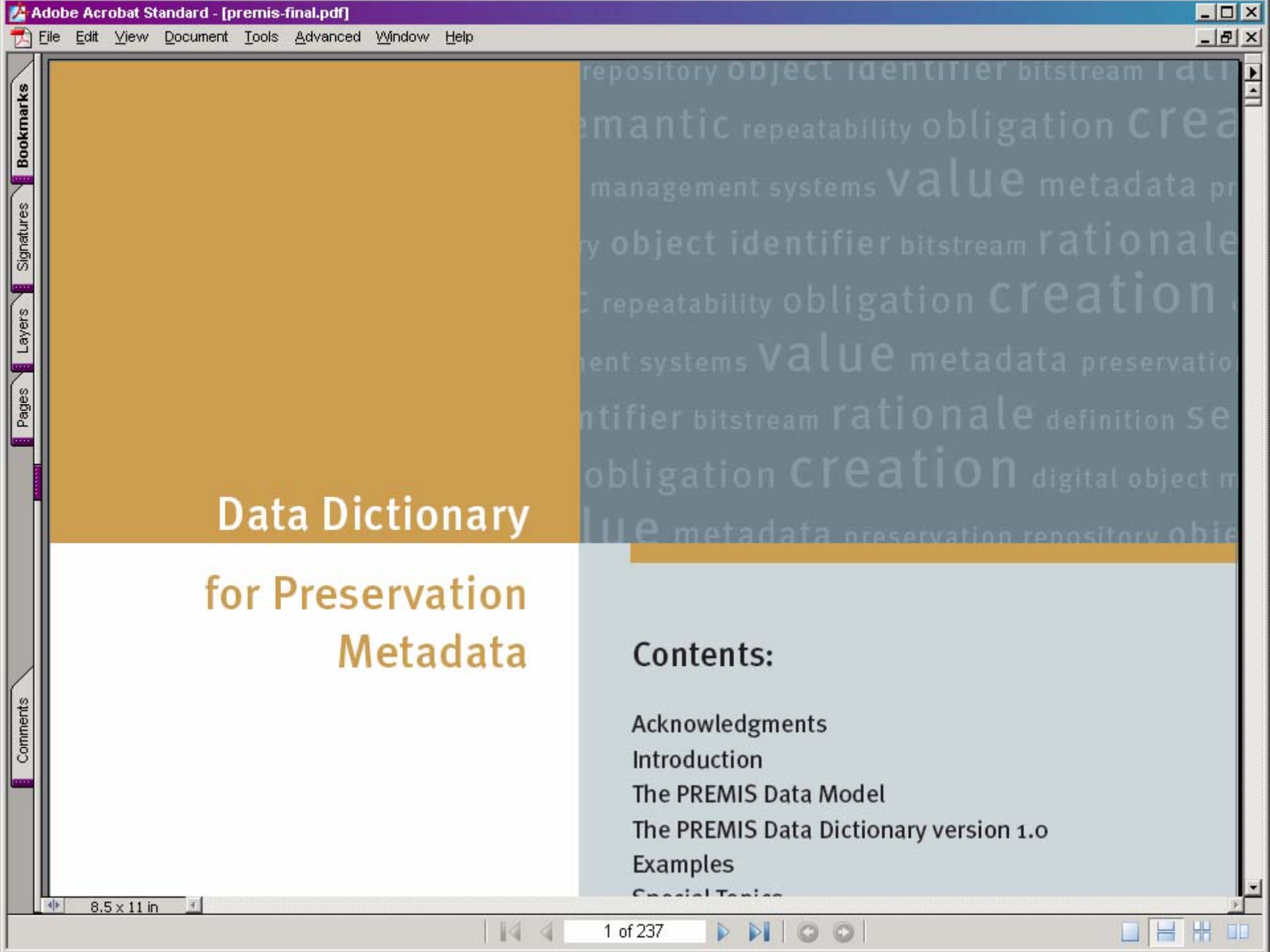
Data Dictionary

for Preservation Metadata

Contents:

PREMIS data dictionary (1)

- – Background:
  - OAIS remains the conceptual foundation (but there are now 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)

- Core preservation metadata:
  - Data Dictionary 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

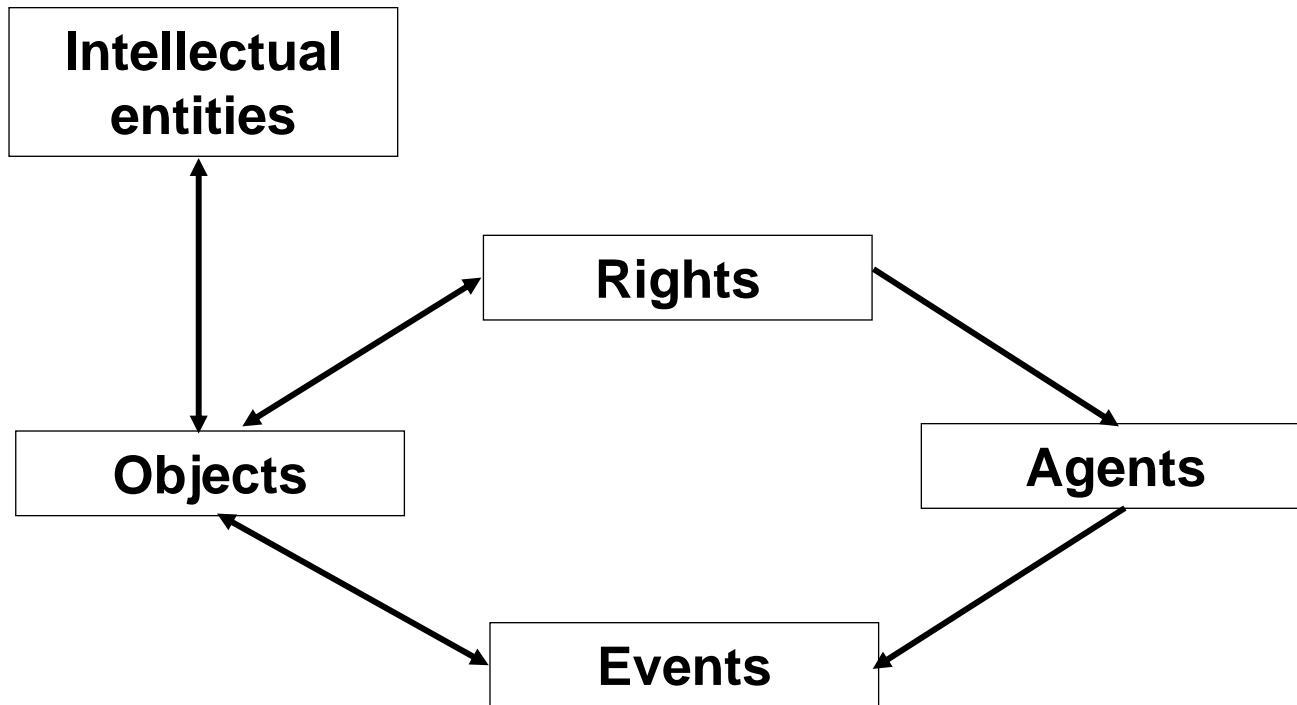- – Based on simple entity-relationship data model that defines five types of entities

PREMIS data model (1)

```
┌─────────────┐
│ Intellectual│
│  entities   │
└──────┬──────┘
       │
       ▼
    ┌───────┐           
    │ Rights│           
    └───────┘           
┌─────────┐        ┌─────────┐
│ Objects │        │ Agents  │
└─────────┘        └─────────┘
       │                │
       ▼                ▼
       ┌─────────────────┐
       │     Events      │
       └─────────────────┘
```

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 concentrates 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

PREMIS documentation

- Data Dictionary, v 1.0
  - Defines semantic units for Objects, Events, Agents and Rights
  - Implementation independent
    - Defines semantics
    - Proposed XML binding

- PREMIS Maintenance Agency
  - Library of Congress
  - http://www.loc.gov/standards/premis/schemas.html

PREMIS limits to scope (1)

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

PREMIS 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 media and hardware 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)

Conclusions

- OAIS is already being used in a variety of contexts:
  - The analysis of existing repository processes
  - Informing the design of systems (and tools)
  - Informing the development of certification criteria
  - The Information Model has influenced the development of preservation metadata standards (e.g. PREMIS) and emerging registries of Representation Information

UKOLN

UNIVERSITY OF BATH

D C C

Key links (1)

- *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1 (2002): http://public.ccsds.org/publications/archive/650x0b1.pdf
- DPC Technology Watch Report on the OAIS model by Brian Lavoie (2004): http://www.dpconline.org/docs/lavoie_OAIS.pdf
- *Assessment of UKDA and TNA Compliance with OAIS and METS standards* by H. Beedham, *et al.*, (2005): http://www.data-archive.ac.uk/news/publications/ oaismets.pdf
- RLG/NARA Task Force on Digital Repository Certification: http://www.rlg.org/en/page.php?Page_ID=580
- CRL Certification of Digital Repositories: http://www.crl.edu/content.asp?l1=13&l2=58&l3=142

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

Preservation of Digital Heritage, Madrid, 14-16 March 2006

Key links (2)

- PREMIS Data Dictionary for Preservation Metadata (2005): http://www.oclc.org/research/projects/pmwg/

- DPC Technology Watch Report on Preservation Metadata by Brian Lavoie and Richard Gartner (2005): http://www.dpconline.org/docs/reports/dpctw05-01.pdf

- DCC Digital Curation Manual Instalment on Metadata by Michael Day (2005): http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/

Muchas gracias por su atención

Thank you for your attention

Acknowledgements

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

Preservation of Digital Heritage, Madrid, 14-16 March 2006