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

Carlo Meghini
Consiglio Nazionale delle Ricerche
Istituto di Scienza e Tecnologie della Informazione
http://nmis.isti.cnr.it/meghini/

CASPAR Training Day for the Cultural and Scientific Domains
January 12, 2009
ICCU (Rome, Italy)

Acknowledgements

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

Session outline

- Background
- Mandatory Responsibilities
- Functional Model
- Information Model

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
- Periodic reviews
- http://public.ccsds.org/publications/archive/650x0b1.pdf

OAIS purpose and scope (1)

- To define an Open Archival Information System (OAIS)
  - An OAIS is 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.
  - The term 'open' means that the document was developed in open forums, and does not imply that access to any OAIS should be unrestricted
  - While an OAIS itself need not be permanent, the information being maintained has been deemed to need "Long Term Preservation"
    - Long term = long enough for there to be a concern about the impact of changing technologies

OAIS purpose and scope (2)

- Primary focus on digital information
  - both as the primary forms of information held and as supporting information for both digitally and physically archived materials.
- The model accommodates information that is inherently non-digital (e.g., a physical sample)
  - but the modeling and preservation of such information is not addressed in detail.

OAIS purpose and scope (3)

- Specific aims include:
  - A framework for the understanding and awareness of the archival concepts needed for long term preservation and access
  - Terminology and concepts for describing and comparing:
    - Architectures and operations
    - Preservation strategies and techniques
    - Data models
  - Consensus on elements and processes for long term preservation and access, and promotes a larger market
  - A foundation for other standards
    - Information NOT in digital form
    - OAIS-related

OAIS purpose and scope (4)

- Applicability:
  - Applicable to any archive, but mainly focused on organisations with responsibility for making information available for the long term
  - Of interest to those who create information that may need Long-Term Preservation and those that may need to acquire information from such archives
  - It does not specify a design or an implementation. Actual implementations may group or break out functionality differently.
- A road map for related standards (section 1.5)

OAIS purpose and scope (5)

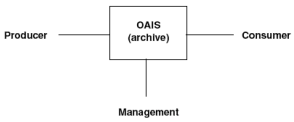
- Conformance
  - An OAIS must support the information model
  - Mandatory responsibilities (section 3.1)
  - The model itself is technology-agnostic
    - "It is assumed that implementers will use this reference model as a guide while developing a specific implementation to provide identified services and content"
    - The model does not assume or endorse any specific computing platform, system environment, system design paradigm, database management system, data definition language, etc.
    - An OAIS may provide additional services
    - A conceptual framework to discuss and compare archives

OAIS high level concepts (1)

- Traditional archives are understood as facilities or organizations which preserve records, for access by public or private communities.
  - The archive accomplishes this task by taking ownership of the records, ensuring that they are understandable to the accessing community, and managing them so as to preserve their information content and authenticity.
- Many other organizations in the government, commercial and non-profit sectors have to take on the information preservation functions because digital information is easily lost or corrupted.

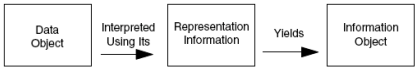
OAIS high level concepts (2)

- OAIS environment
  - Producer provides the information
  - Management sets overall policy (not the day-to-day operations)
  - Consumer finds and acquires preserved information of interest
    - Designated Community is the set of Consumers for whom the information is preserved and who should be able to understand the preserved information.



OAIS high level concepts (3)

- A person, or system, can be said to have a Knowledge Base, which allows them to understand received information.
- Information is any type of knowledge that can be exchanged, and is expressed by some type of data.
  - The information in a book is typically expressed by characters (the data) which, when combined with a knowledge of the language used (the Knowledge Base), are converted to more meaningful information. If the recipient does not know the language, then the book needs to be accompanied by dictionary and grammar (i.e., **Representation Information**) in a form that is understandable using the recipient's Knowledge Base



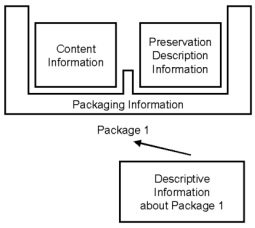
OAIS high-level concepts (4)

- In order for this Information Object to be successfully preserved, it is critical for an OAIS to clearly identify and understand the Data Object and its associated Representation Information.
  - For digital information, this means the OAIS must clearly identify the bits and the Representation Information that applies to those bits.
- The OAIS must understand the Knowledge Base of its Designated Community to understand the minimum Representation Information that must be maintained.

OAIS high-level concepts (5)

- The unit of exchange between an OAIS and its surrounding the environment is an Information Package.
- An Information Package is a conceptual container of two types of information:
  - Content Information and
  - Preservation Description Information (PDI).
- An information package is discoverable by virtue of the Descriptive Information

OAIS high level concepts (6)



Information Package Concepts and Relationships (Figure 2-3)

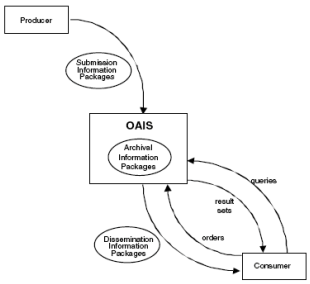
OAIS high-level concepts (7)

- The Packaging Information is that information which, either actually or logically, binds, identifies and relates the Content Information and PDI.
- The Descriptive Information is that information which is used to discover which package has the Content Information of interest.

OAIS high-level concepts (8)

- Information Package variants
  - Submission Information Package (SIP)
  - Archival Information Package (AIP)
  - Dissemination Information Package (DIP)
- Packages will need to vary depending upon their role
  - For example, imaging and e-journal projects often differentiate between their well-managed (and described) "master" files and the derived versions (thumbnails, JPEG files, PDFs) made available through the Web

OAIS external interactions (1)



OAIS external interactions (2)

- High level view of the interactions in an OAIS environment
  - Management interaction
    - Charter and scope, Funding, Evaluation, Conflict resolution
  - Producer interaction
    - Submission agreements
  - Consumer interaction
    - Help desk questions, information discovery (on Description Information), ordering of information

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.
  - the community should understand the information without the assistance of the experts who produced the information
- Follow documented policies and procedures which

1. ensure that the information is preserved against all reasonable contingencies, and
2. enable the information to be disseminated as authenticated copies of the original, or as traceable to the original

OAIS mandatory responsibilities (3)

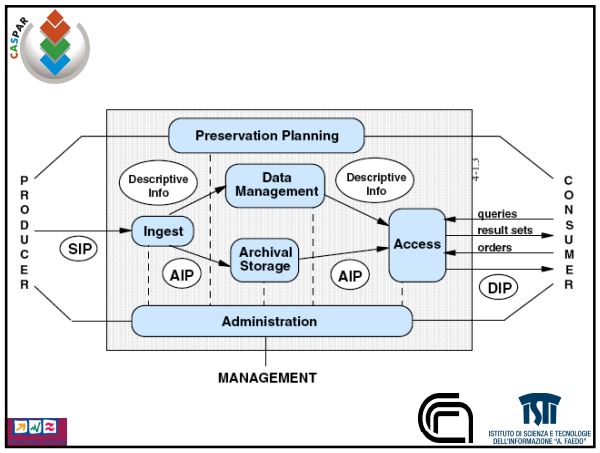
- Make the preserved information available to the Designated Community
- Section 3.2 exemplifies mechanisms for discharging responsibilities

OAIS Functional Model

(Section 4.1)

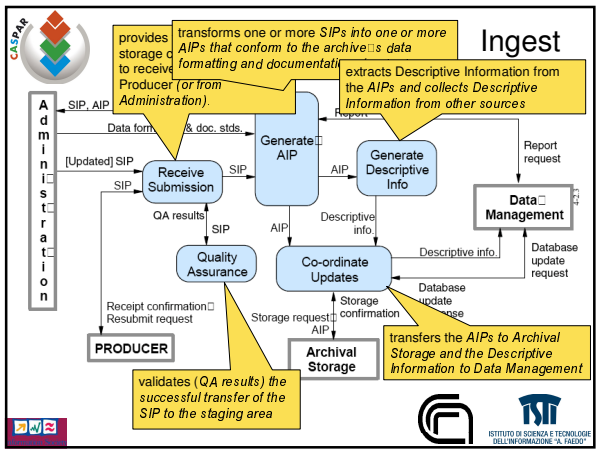
OAIS Functional Model (1)

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



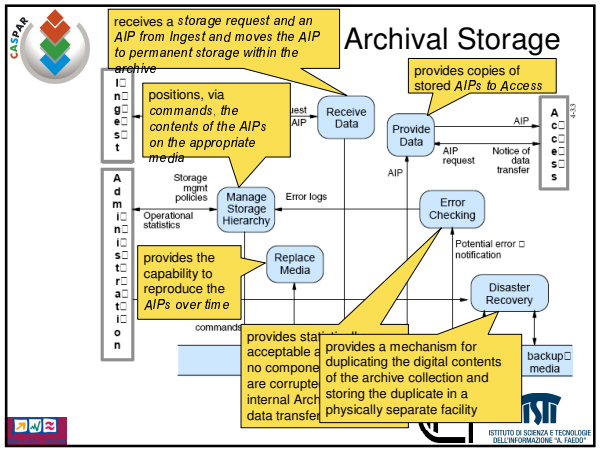
Ingest

Provides the services and functions to accept Submission Information Packages (SIPs) from Producers (or from internal elements under Administration control) and prepare the contents for storage and management within the archive.



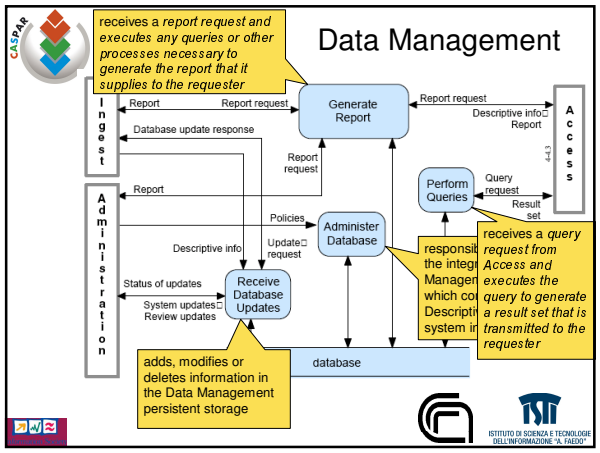
Archival Storage

Provides the services and functions for the storage, maintenance and retrieval of AIPs.



Data Management

Provides the services and functions for populating, maintaining, and accessing both Descriptive Information which identifies and documents archive holdings and administrative data used to manage the archive.



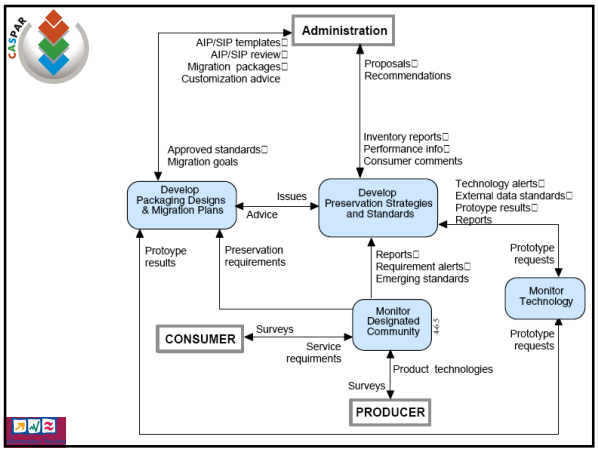
Administration

Provides the services and functions for the overall operation of the archive system, including:

- soliciting and negotiating submission agreements
- auditing submissions to ensure that they meet archive standards, and
- maintaining configuration management of system hardware and software.

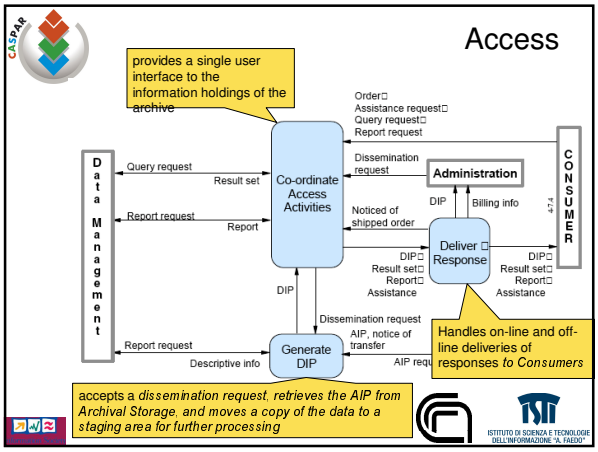
Preservation Planning

Provides the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible to the Designated User Community over the long term, even if the original computing environment becomes obsolete.



Access

Provides the services and functions that support Consumers in determining the existence, description, location and availability of information stored in the OAIS, and allowing Consumers to request and receive information products.



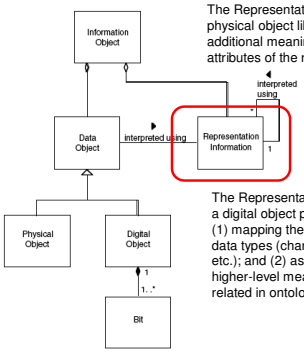
OAIS Information Model

(Section 4.2)

Background

- The primary goal of an OAIS is to preserve information for a designated community over an indefinite period of time.
- To this end,an OAIS must store significantly more than the contents of the object it is expected to preserve.
- The information model describes the types of information that are exchanged and managed within the OAIS .

OAIS Information Object



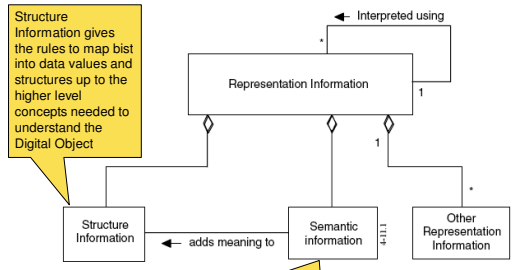
The Representation Information accompanying a physical object like a moon rock may give additional meaning to the physically observable attributes of the rock.

The Representation Information accompanying a digital object provides additional meaning by (1) mapping the bits into commonly recognized data types (character, integer, strings, records, etc.); and (2) associating these data types with higher-level meanings that are defined and inter-related in ontologies.

Representation Information

The purpose of the Representation Information object is to convert the bit sequences into more meaningful information

Structure Information gives the rules to map bist into data values and structures up to the higher level concepts needed to understand the Digital Object

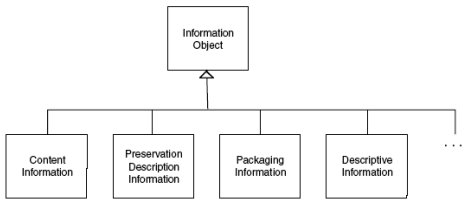


the information needed to make the Digital Object understandable by the Designated Community

Representation Information Networks

- Representation Information may contain references to other Representation Information
- Representation Information is itself an Information Object that may have its own Digital Object and other Representation Information for understanding the Digital Object
- The resulting set of objects can be referred to as a **Representation Network**.

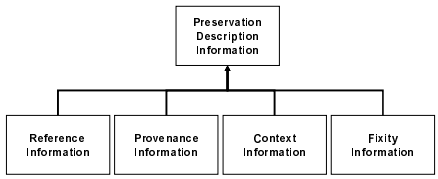
Types of information objects



Content Information

- The Content Information is the set of information that is the original target of preservation by the OAIS.
- The Content Information is the Content Data Object together with its Representation Information. The Content Data Object in the Content Information may be either a Digital Object or a Physical Object (e.g., a physical sample, microfilm).
- Any Information Object may serve as Content Information.

Preservation Description Information



PDI Preservation Description Information (Figure 4-16)

Preservation Description Information

- **Reference Information**: identifies and describes one or more mechanisms used to provide assigned identifiers for the Content Information. It also provides those identifiers.
- **Context Information**: documents the relationships of the Content Information to its environment (why the Content Information was created and how it relates to other Content Information).

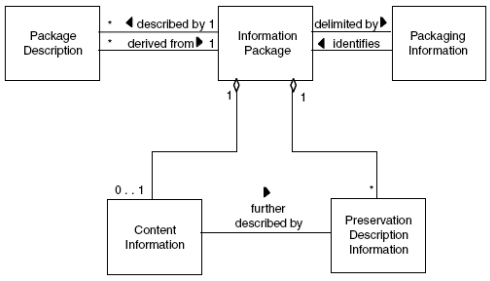
Preservation Description Information

- **Provenance Information**: documents the history of the Content Information (origin or source, changes and custody) Provenance can be viewed as a special type of context information.
- **Fixity Information**: provides the Data Integrity checks or Validation/Verification keys used to ensure that the particular Content Information object has not been altered in an undocumented manner.

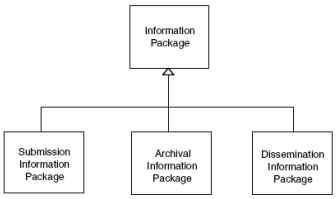
OAIS Information Packages

- The conceptual information structures required to accomplish the OAIS functions.
- An Information Package is a container.
- There are several types of Information Packages that are used within the archival process. These Information Packages may be used
  - to structure and store the OAIS holdings (AIP);
  - to transport the information from the Producer to the OAIS (SIP)
  - to transport requested information between the OAIS and Consumers (DIP).

OAIS Information Package



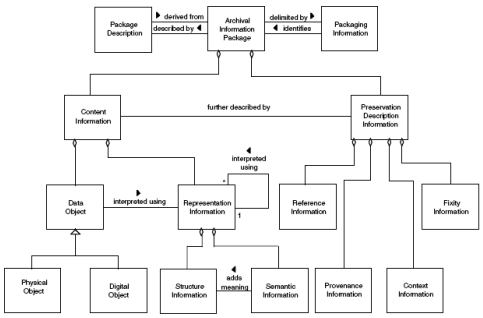
Information Package Types



SIP

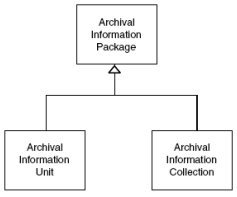
- The form and detailed content of a SIP are typically negotiated between the Producer and the OAIS.
- Most SIPs will have some Content Information and some PDI, but it may require several SIPs to provide a complete set of Content Information and associated PDI.
- If there are multiple SIPs that use the same RepInfo, it is likely that such RepInfo will only be provided once.
- Within the OAIS, one or more SIPs are transformed into one or more AIPs for preservation.

AIP



Types of AIPs

An AIU is viewed as having a single content Information Object that is described by exactly one set of PDI.



An AIC Content Information is viewed as a collection of other AICs and AIUs, each of which has its own PDI. In addition, the AIC has its own PDI that describes the collection criteria and process.

DIP

- In response to an Order, the OAIS provides all or a part of an AIP to a Consumer in the form of a DIP.
- The DIP may also include collections of AIPs, depending on the dissemination agreement betwen OAIS and Consumer.
- The Packaging Information will always be present so that the Consumer can clearly distinguish the information ordered.
- The purpose of the Descriptive Information of a DIP is to give the Consumer enough information to recognize the DIP from among possible similar packages.

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)

References

- *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
- Trusted Repositories Audit & Certification: http://www.crl.edu/PDF/trac.pdf

Implementing the OAIS model

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 formal issue by ISO
- Main areas of influence:
  - Related standards (e.g., CCSDS Archive-Producer Interface)
  - Standardising terminology
  - Compliance and certification
  - Analysis and comparison of archives
  - Informing system design
  - Preservation metadata

Compliance and certification

OAIS compliance (1)

- Many repositories or preservation tools claim OAIS influence or compliance:
  - e.g., IBM DIAS, DSpace, OCLC Digital Archive, METS, the list is endless
  - LOCKSS System has produced a "formal statement of conformance to ISO 14721:2003" (lockss.stanford.edu/)
- The OAIS model's own view (OAIS 1.4):
  - Supporting the information model (OAIS 2.2),
  - Fulfilling the six mandatory responsibilities (OAIS 3.1)

OAIS compliance (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
  - Built upon the OAIS model …

Trusted digital repositories (2)

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

Digital repository certification (1)

- RLG-NARA Task Force on Digital Repository Certification
  - RLG and the US National Archives and Records Administration
  - 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)
- Various certification initiatives (CRL, DCC, nestor, DRAMBORA)

Digital repository certification (2)

- Trusted Repositories Audit & Certification (TRAC): Criteria and Checklist (March 2007)
  - Organisational infrastructure
    - e.g., governance, organisational structures, mandates, policy frameworks, funding systems, contracts and licenses
  - Digital Object Management (OAIS functions)
    - e.g., ingest, metadata, preservation strategies
  - Technologies, Technical Infrastructure, & Security

The analysis and comparison of repositories

The analysis of existing services

- A process that was 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

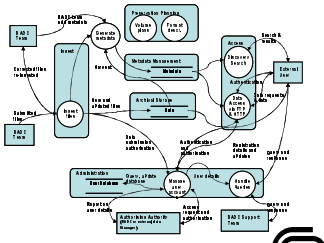
BADC/APS case study

- 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
  - Revealed that AIP needed better definition
  - Weaknesses identified with the Preservation Planning role, e.g. little explicit monitoring of technology or of 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)

- Project funded by the UK Joint Information Systems Committee (JISC)
- Partners:
  - UK Data Archive
  - The National Archives
- Aimed to map UKDA and TNA to OAIS functional and information models, a "use case" for compliance
- Beedham, *et al.*, *Assessment of UKDA and TNA Compliance with OAIS and METS Standards* (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 six OAIS Mandatory Responsibilities are carried out by almost any well-established archive
  - The OAIS Designated Community concept assumes a identifiable and relatively homogenous user community; this was not the case for either UKDA or TNA
  - The relationship between AIPs and DIPs needed clarification

UKDA and TNA case study (3)

- Conclusions (continued):
  - 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)
  - But ... The use of 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 generally 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:
    - IBM DIAS (used by Koninklijke Bibliotheek)
    - OCLC Digital Archive Service
    - aDORe (Los Alamos National Laboratory)
    - Stanford Digital Repository
    - MathArc (Cornell UL and SUB Göttingen)
  - Tools:
    - Repository software: DSpace, FEDORA, …
    - DCC Representation Information Repository and 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/

Informing system design (4)

- Research projects
  - OAIS is the “guiding principle” of CASPAR
  - CASPAR Conceptual model
  - Representation Information registries and repositories

Preservation metadata

Preservation metadata

- 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)
  - Will be dealt with in more detail in a separate session

Summary

Summary

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

Ingest exercise

Ingest exercise (1)

- Select a scenario, e.g.:
  - National library building a collection of e-journals
  - University library setting up an institutional repository to collect e-prints produced by academic staff
  - Museum or archive digitising photographic images
  - ...
- Your director has asked you whether your service conforms to the OAIS standard
- You are now looking in detail at your repository processes and policies and are evaluating how they relate to OAIS terms and concepts

Ingest exercise (2)

- For this exercise, we will only consider the Ingest function
  - Ingest is understood as those 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
- Main functions:
  - Pre-Ingest - negotiation and agreement on the nature of SIPs
  - Receive Submission
  - Quality Assurance - for successful transfer
  - Generate AIP - the version stored in Archival Storage
  - Generate Descriptive Information - could be extracted from AIPs
  - Co-ordinate Updates - transfers AIP to Archival Storage and Descriptive Information to Data Management

Ingest exercise (3)

- Think about requirements for defining SIPs and generating AIPs
  - Remember that Information Packages are more than just the content itself - also includes some level of Representation Information
  - Ingest is the main interface between the OAIS and the Producers of content
    - Producers will have their own requirements
    - The level of "control" over Producers will vary, depending on context
  - The OAIS needs to make decisions on:
    - What it can accept (the SIP)
    - Its own requirements for the stored version (the AIP)
    - How to generate an AIP

Ingest exercise (4)

- Things to consider for your scenario:
  - What type of objects are you receiving?
  - How will you receive them?
  - What formats are involved?
  - What level of control do you have over the Producer(s)?
  - What are your main requirements for an AIP? (significant properties)
  - What Representation Information will you need?
  - What other types of metadata (Preservation Description Information, Descriptive Information) will you need?
  - Can the Producers supply any of this metadata? If so, how?
  - How will you package content and metadata in Information Packages?

Feedback and discussion

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 Joint Information Systems Committee and the UK Research Councils' e-Science Core Programme: http://www.dcc.ac.uk/