DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

Preservation & Curation as a Digital Library Challenge: From the Perspective of the DELOS_NOE Preservation Cluster

ICA2004, Vienna
DELOS Workshop on Preservation in Digital Libraries
Tuesday, 24th August 2004 Room A353

Dr Seamus Ross
Professor of Humanities Informatics and Digital Curation
Director, HATII(University of Glasgow)

Information Society

DELOS Workshop

- Introductory Presentation on DELOS and the Preservation Challenge
- Examination of File Formats and OAIS
  - Robert Sharpe of Tessella
- Break
- Introduction to Repository Models
  - James Currall of Glasgow University
- Round-table discussion

Objective of digital preservation

"retaining the ability to display, retrieve, manipulate, and use digital information in the face of constantly changing technology"

Making the Information & Knowledge Environment Work

If the national and global network infrastructures are to provide a suitable business environment then systems must be in place to guarantee that the requirements for: integrity, authenticity, reliability and the archiving of digital information can be carried out easily and effectively

Key Preservation Issues

- Medium
  - **storage media naturally decay**
- Technological (e.g. hardware/software)
  - hardware and software **obsolescence makes data/information inaccessible**
- Intellectual
  - **validation of integrity and authenticity**
- Contextual
  - **avoid loss of meaning with metadata**
- Legal Impediments
- The Organisation and its staff

Digital Library Service Phases

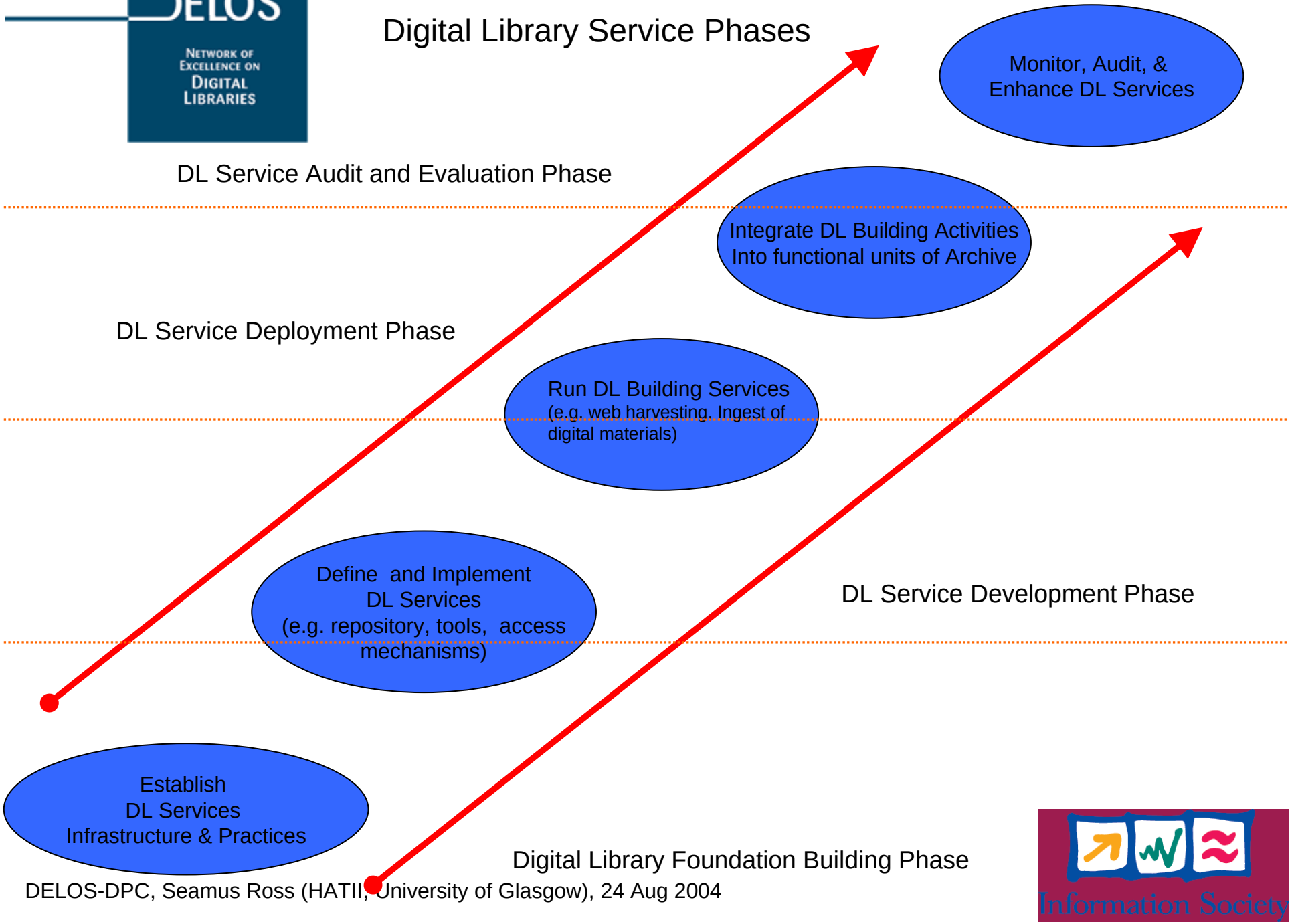
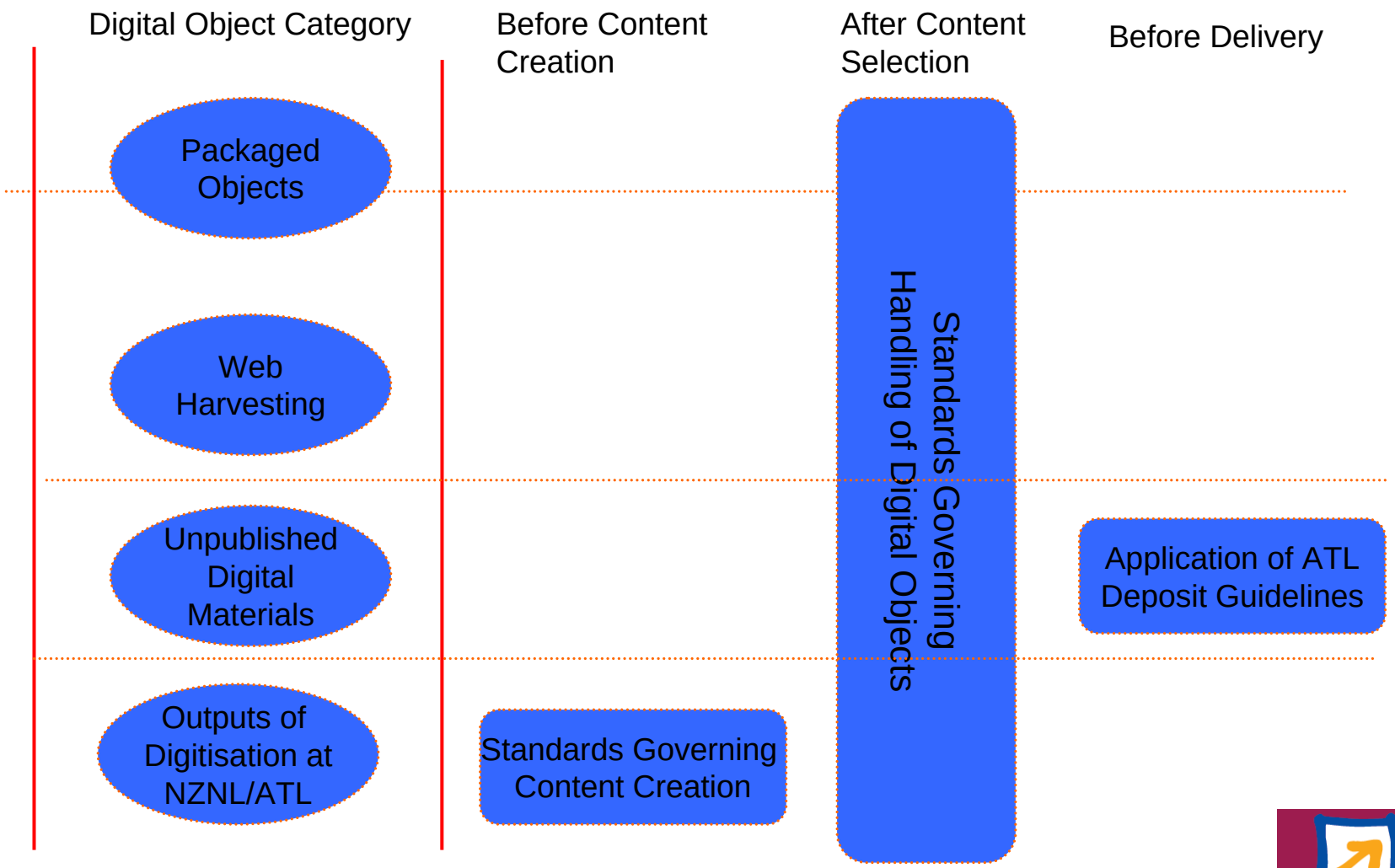


Diagram 1: Place and Role of Standards



What's Happening in Preservation

- Collaborative Projects—
  - From Pittsburgh, InterPARES, NEDLIB, to CEDARS/CAMiLEON, Presto
- National Initiatives led by Libraries & Archives
  - Such as ….. UK (NDAD), NARA, BL, Library of Congress, BNF, Danish Web Archiving Project
- Gap between commercial activity and the knowledge in the public sector about these
- Legal Challenges to digital preservation (IPR, privacy) but FOI & e-citizenship (or e-government) may provide counterbalances
- Recognition that our cultural memory is at risk and that it is composed of many types of digital objects (e.g. audio, VR)
- Institutional Missions often without synergy of effort
- OAIS
- Trusted Repositories

Appraisal and Selection

- **Administrative value**
- **Evidential value**
  - **(e.g. product liability)**
- **Informational value**
- **Reusability & Integration**
- **Technical viability**
- **Anticipated costs of preservation**
- **Usage restrictions**

RM involvement in system design stage essential

Access, Intelligibility, and Maintainability

- Topology of data/information resources
- Hardware & software issues
- Migration and selection
- Storage strategies
- Migration and preservation infrastructure

Materials must be identified for preservation before they are created if activities, processes and systems are to support their preservation.

Recurring Value of Digital Objects

- **Industry dependent**
- **Product liability**
- **Competitive advantage**
- **Recurring value through reuse**
- **Commercially valuable information a candidate for preservation**
- **Corporate memory**
- **Costs of re-creation vs storage**
- **Foundation for scholarly endeavour**

Information Risks

- Uncontrolled growth in data and records
- Possibility of accidental record loss
- Security (e.g. information leakage)
- Record duplication and authentication
- Unauthorised modification of records
- Loss of integrity and authenticity of digital resources

Preservation Options

- **Hardware and software preservation**
  - technically complex and expensive
- **Software & Hardware emulation**
  - practical (?)
- **Data migration**
  - can lead to data and information loss
  - can lead to loss of functionality
- **Virtual Machines**
- **Binary Retargetable Code**
  - Transmogrifying Adaptable Preservation (TAP)

Digital Archaeology

- It is very difficult to lose your data completely, although it may be very expensive to recover it. So act to avoid compromising your valuable resources, but don't panic when all goes wrong because everything need not be lost.

DELOS: The Main Objective

To define and conduct a joint program of activities in order to integrate and coordinate the on-going research activities of the major European research teams in the field of digital libraries for the purpose of developing the next generation digital library Technologies – http://www.delos.info

Other Objectives

- Network and structure European research on digital libraries, so as to consolidate an emerging community
- Contribute towards improving the effectiveness of European research in the digital library field
- Provide a forum where researchers, practitioners, and representatives of interested applications and industries can exchange ideas and experiences
- Promote cooperation between European and national digital library initiatives
- Improve international cooperation in DL research areas

Activity Organisation

- The Network activities are organised into a set of Workpackages/clusters:
  - WP1: Digital Library Architecture
  - WP2: Information Access and Personalization
  - WP3: Audio/Visual and Non-Traditional Objects
  - WP4: User Interfaces and Visualization
  - WP5: Knowledge Extraction and Semantic Interoperability
  - WP6: Preservation
  - WP7: Evaluation
  - WP8: Dissemination and Spreading of Excellence

Digital Library Architecture

Initial Focus--

- To evaluate, from the conceptual and experimental point of view, the three main directions to a Digital Library architecture:
  - Service based Architecture
  - P2P Architecture
  - Grid based Architecture

Initial Focus--

- To establish a common foundation for European researchers in all areas related to Information Access and Personalization in Digital Libraries, including accessing information in single sources, integrating information from multiple sources, and personalization.

Audio/Visual and Non-Traditional Objects

Initial Focus--

- To establish a common ground of knowledge for European researchers about emerging requirements and research directions and solutions for standards, management, access, presentation and use of digital libraries content and make available testbeds and demonstrators of the most significant solutions.
- To advance the state-of-the-art in the solutions for metadata extraction, object management, information access and content-based retrieval, taking into account the specific requirements of the delivery media, the application requirements and the user terminals.

User Interfaces and Visualization

Initial Focus–

- To elaborate a common understanding of the role and scope of user interface research in the digital library area.

- To develop the theoretical framework for digital library user interface design

Knowledge Extraction and Semantic Interoperability

Initial Focus--

- To explore the potential of new models, algorithms, methodologies and processes in a variety of technical applications, institutional frameworks and cross-sectoral environments.
- To create guidelines and recommendations of best practice for dissemination to the widest possible community of interest.

Evaluation

Initial Focus--

- To enable discussions between evaluation specialists and developers and spreading evaluation knowledge among the latter.
- To develop new evaluation toolkits for implementing these methods.
- To continue existing evaluation initiatives, with a special focus on those aspects which are relevant to the digital library application area.

Preservation

Initial Focus–

- To catalyse the research and funding environment to enable of delivery of the DELOS/NSF research agenda for Digital Preservation and Archiving.
- To lay the foundation for testbeds and necessary metrics and tools for assessing preservation strategies.
- To ensure access to file format information and to establish the relationship between a typology of file formats and preservation strategies.
- To ensure that system development methodologies reflect preservation analysis and design issues.

The Preservation Cluster

- catalyse the research and funding environment to enable of delivery of the DELOS/NSF research agenda for Digital Preservation and Archiving,
- to lay the foundation for testbeds and necessary metrics and tools for assessing preservation strategies,
- to raise the profile of digital preservation issues within the Digital Library Community,
- to collaborate with other international bodies to ensure consistencies of digital repository standards,
- to ensure access to file format information and to establish the relationship between a typology of file formats and preservation strategies,
- to enable the definition of attributes and functionalities that need to be represented, and
- To ensure that system development methodologies reflect preservation analysis and design issues.

The Preservation Cluster

- **Task 0: Cluster Management**
  - ***Task Leader and Contact Person: Seamus Ross, UG (9); Participants: UG (9)***
  - **This task will oversee the work of the Preservation cluster. It will include the following activities:**
    - **organisation of cluster workshops in which the past and future activities of the cluster will be discussed and monitored**
    - **setting up and maintenance of the cluster website**

The Preservation Cluster

- **Task 1: Digital Preservation Testbed Forum.**
  - ***Task Leader: Hans Hofman (NANETH)***
  - ***Participants: TUW , UCO, NANETH, UG, UKOLN***
- **Establish a framework of a digital preservation testbed environment.**
- **Produce metrics for testing and validating digital preservation strategies.**
- **Establish mechanisms for ensuring comparability between testbed environment including a testbed test data set (which might include programmes as well ad data).**
- **Delivery month 12**

The Preservation Cluster

- **Task 2: Designing, Deploying and Managing Digital Repositories.**
  - ***Task Leader and Contact Person: Seamus Ross, UG (9)***
  - ***Participants: UG (9), OEAW (37), NANETH (35), UNIURB (30), UCO (39)***
- **Contribute to the development of digital repository frameworks and mechanisms for validating the suitability of digital repository implementations.**
- **Evaluate the current and emerging systems and storage models for digital repositories.**
- **Delivery month 12**

The Preservation Cluster

- **Task 3: File Formats, Classification, and Typology**
  - ***Task Leader and Contact Person: Maria Guercio, UNIURB (30)***
  - ***Participants: UNIURB (30), UG (9), OEAW (37), UKOLN (4), TUW (29)***
- **Contribute to the development of file format registries and the mechanisms for their use.**
- **Define the relationship between file format types and preservation methods and to assess the viability of producing generic metrics to measure the viability of this preservation approach.**
- **Define a typological framework and the rules for extending the framework.**
- **Delivery date: month 12**

The Preservation Cluster

- **Task 4: Documentation of Functionality and Behaviour Metrics.**
  - ***Task Leader and Contact Person: Andreas Rauber TUW***
  - ***Participants: UG (9), UCO(39), TUW(29), NANETH (35),***
- **Define framework for documenting behaviour and functionality.**
- **Develop an overview of the attributes of functionality and behaviour that need to be represented and mechanisms for representing them.**
- **Establish the viability of automating the process of functionality and behaviour verification.**
- **Delivery date: month 18**

The Preservation Cluster

- **Task 5: Enabling the Integration of Digital Preservation Architectures:**
  - ***Task Leader and Contact Person: Manfred Thaller, UCO (39)***
  - ***Participants: UCO (39), TUW (29), UKOLN (4), UofGlagow(9)***
- **Develop the requirements for a preservation functionality modelling tool.**
- **Develop a method for modelling preservation functionality that could be integrated with more traditional system design and development methodologies (SSADM).**
- **Encourage take up testing of the tool through presentations at software engineering conferences.**
- **Delivery date: month 18**

DELOS-DPC only a Start

- A narrow range of research questions
- There are many more that need to be addressed
- See NSF/DELOS Working Group on Digital Preservation

Difficulties Facing Creators and Users

- What information should be retained?
- Where should it be stored?
- What about the diversity of document types?
- How do I access it if I need it?
- How long should it be kept?
- What is its value?
- What are the costs and justifications?
- Does record creation equal retention?

Preservation questions

- What should be archived?
- What levels of documentation will be required?
- What selection criteria should be used?
- What standards should be used?
- Who pays? Who uses? Who selects?
- Medium, environment, context, integrity

Organisational Obstacles to Preservation

- Tendency towards decentralisation & networked organisational structures
- Lack of collaboration between records managers, creators, and IT staff
- Need to link records management strategies with organisational objectives
- Lack of organizational commitment (social, economic, political)
- Failure to acknowledge the investment needed
- Failure to identify recognizable benefits
- Failure to link Preservation to corporate objs.

Migration strategies

- Sequence of tasks undertaken periodically
- Change media
- Converting format or encapsulating
- Incorporation standards
- Time & labour dependencies
- Costs vs. value
- Influenced by processes, systems, & best practice

El Archivo General de Indias

Preparation strategies

- Design systems for long term accessibility (e.g work with developers)
- Control, monitor, document, and audit migration
- Avoid proprietary systems (e.g. hardware, software, applications, standards)
- Avoid emerging technologies

Documentation & Metadata

- Information identifying the resource
- Terms of access
- Guidelines to open and read
- Details of how, when and why the record
- Clues to its authenticity and verification
- Evidence of its use
- Assistance with the meaning of the record
- Essentially: structure, context, [content], use history

System Documentation: Metadata Elements

- Logical and physical models of the system
- Information flow models
- Data flow diagrams
- Entity-relationship charts
- Process model descriptions
- Data dictionaries
- Information Resource Directory Systems

Preservation Metadata

- Comprehensive metadata framework applicable to the digital preservation activity
  - RLG/OCLC Working Group on Preservation Metadata
  - New Zealand National Library Metadata Framework

Metadata Consensus Gaps

- Testing of model across organisational types
- Representation of business processes (including information flow)
- Layer of documentation to cover system documentation
- Incorporation of metadata guidelines into software

OAIS Model

- **OAIS = Open Archival Information Systems**
- **Key players in development**
  - **National Space Science Data Centre**
  - **Consultative Committee for Space Data Systems**
- **Premises Underlying OAIS**
  - **Data are irreplaceable (esp observation)**
  - **Data and associated metadata must be moved across technologies**
  - **Representations and formats will change**
  - **Lack of consensus on adequate metadata standards**

Key OAIS Objectives

- **Objective**
  - recognised no framework for developing digital archive standards
  - need for a reference model
  - recognise the hybrid nature of archives
  - collaborate with archival community
  - focus on data resulting from space missions
  - near-term and indefinite storage of digital data
  - independent of implementation model
  - address full range of archival processes

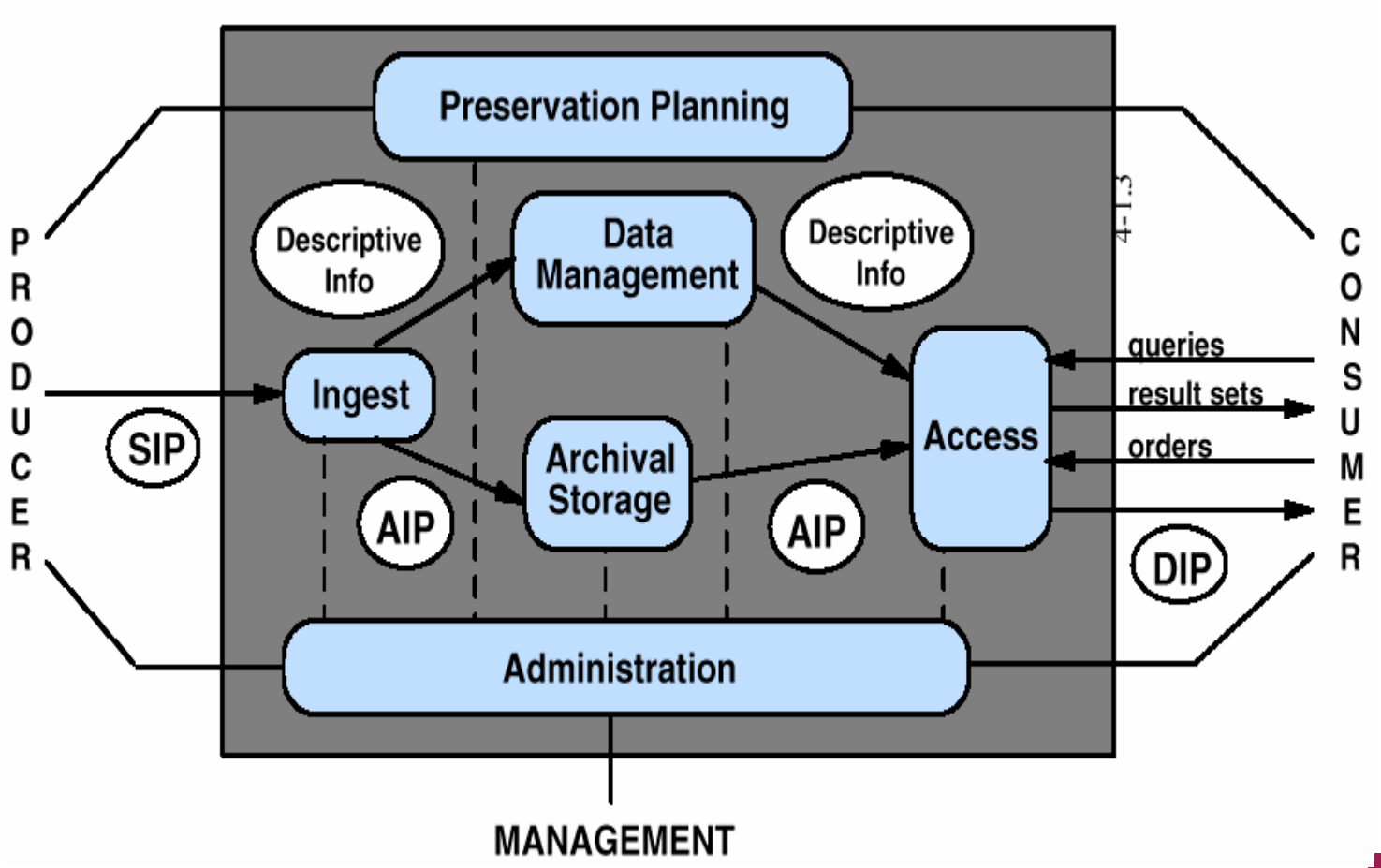
OAIS Overview

- Manages ingest of Information Packages from creators
- Defines the communities needing the Information
- Reflects needs of identified user community
- Enables preservation in an understandable way
- Uses documented policies and procedures

Advantages of OAIS

- Provides a model where one was lacking
- Facilitates procurement of systems
- Enables interoperability between OAIS complaint systems
- Supports the migration task
- Lays out a minimum set of responsibilities

OAIS MODEL


Preservation Planning
Descriptive Info
Data Management
Descriptive Info
PRODUCER
SIP
Ingest
Archival Storage
Access
AIP
AIP
queries
result sets
orders
DIP
CONSUMER
Administration
MANAGEMENT
4-1-3

Who is working with OAIS

- Archive & Library Community
  - Koninklijke Bibliotheek (KB) through NEDLIB—design and architecture of Deposit System for Electronic publications
  - CEDARS
  - NARA and the San Diego SuperComputer Center
  - National Space Science Data Center
  - Pharmaceutical & Aerospace Industries
  - French Space Agency for its plasma physics archive

Trusted Repositories

- What is one?
- RLG/OCLC Proposal
  - need a programme for certifying trusted repositories
  - checklist of concept and key elements needed
- Depends on definable, certified and auditable practices
- What would certification guarantee and how would it be revoked and with what implications

Aspects need certification

- People
  - through developing competencies
- data
  - Quality management, policy, validation
- processes
  - OAIS model, IPR, FOI, organisational practices
- managing organisations
  - audit of approaches organisations take to data management

Certification

- Statement of attributes to be measured
- Policies and Assumptions (e.g. practices, environment and security)
- Procedures against standards
- Relationship with depositors
- What processes are in place to manage fidelity checks for ingest
- What metadata processes are in place
- What user needs evaluation work is carried out

Organisations Need Help

- Off-the-shelf policy statements
- Business cases & strategies
- Digestible guidance on technologies and their preservation implications
- Improved models (reference, costs, standards, functional requirements)
- Simple Guidelines on digital survival
- Access to Metadata Repositories
- Guidance on creating data repositories
- IPR support and guidance

Thinking Toward the Future

- Design information for long term accessibility (e.g work with creators)
- Migration must be controlled, monitored, documented, and audited
- Avoid proprietary systems and emerging technologies
- better software
- intelligent record selection & appraisal tools
- mechanisms for maintaining links between business process and records created/used by them
- case studies of the cost-benefits analysis for data loss vs preservation

Acquisition of Experience

- Develop test experimental frameworks
- Experiment with ingesting, managing, and providing access digital assets
  - Netherlands: Digital Repository Project
  - US: NARA
- Do something concrete -- gain experience
- Ensure parameters of the research are well-documented so that they can be duplicated
- Aim for ‘recipe-like’ descriptions of processes

DELOS -DPC Summer School

- **Digital Preservation Summer School**
  - **5 - 11 June 2005**
  - **Co-Directors Hans Hofman (Dutch National Archives) & Seamus Ross (HATII)**
  - **Southern France**
  - **Lectures by a dozen leading practitioners**
  - **Official Announcement: 21 October 2004**