Preservation Metadata Initiatives: Practicality, Sustainability, and Interoperability

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

ERPANET Training Seminar: Metadata in Digital Preservation, Archivschule Marburg, Germany
3-5 September 2003

Presentation outline

- Categorisation of standards
- Practical issues
  - Implementation
  - Sustainability
- Interoperability
  - Registries of formats and metadata

Basics

- Digital preservation strategies 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)
  - "It's all about metadata" (Kelly Russell, ca. 2000)
- Metadata fulfil various roles, e.g.:
  - Within a digital repository, “metadata accompanies and makes reference to each digital object and provides associated descriptive, structural, administrative, rights management, and other kinds of information” (Clifford Lynch, 1999)

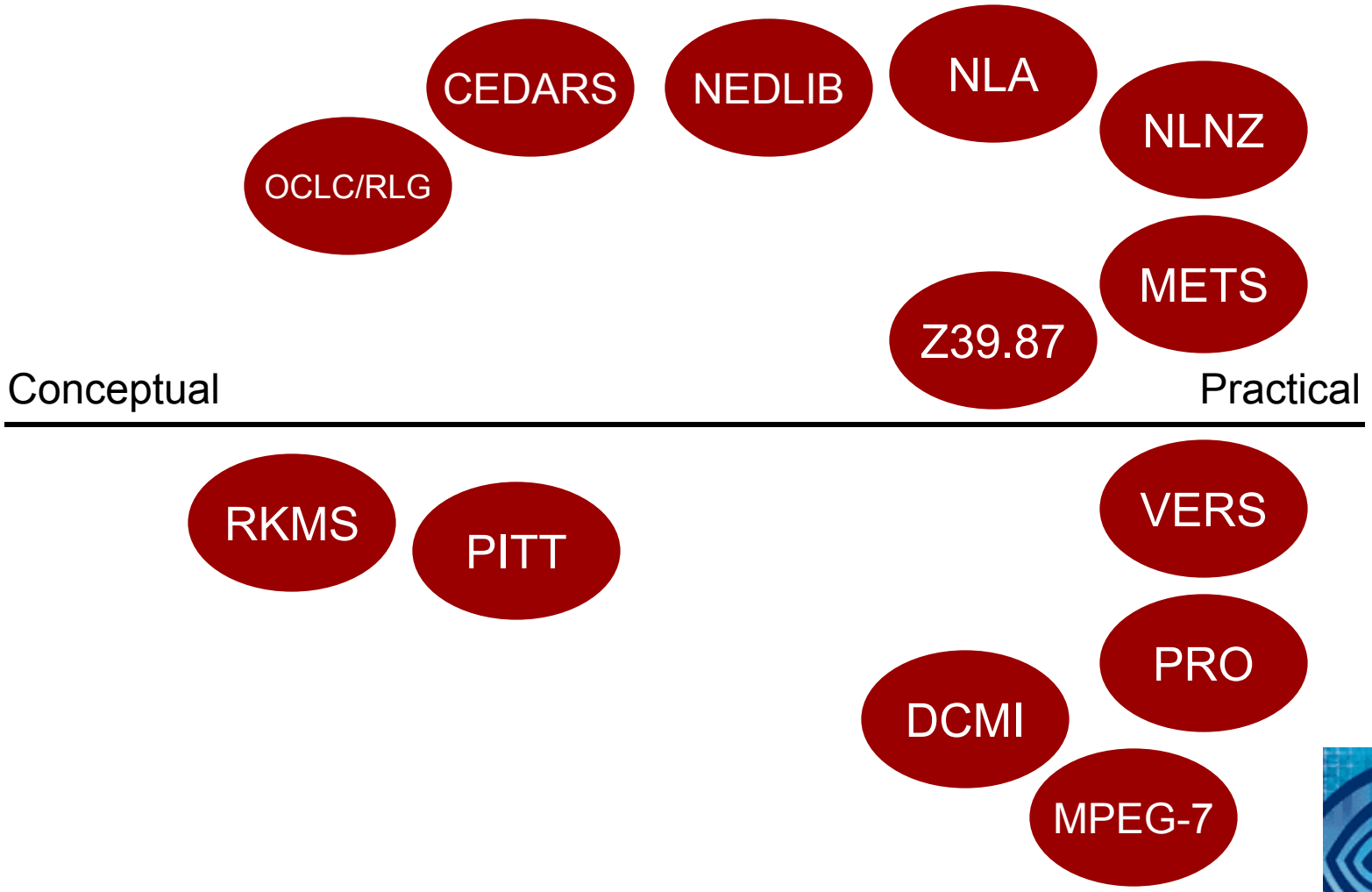
The OAIS model

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

Existing standards

- Developed from many different perspectives:
  - Generic
    - Applications of DCMES
  - Digital libraries:
    - OCLC/RLG Framework, Cedars, NEDLIB, NLA, NLNZ, METS, NISO Z39.87 …
    - OAIS influence has been greatest in this area
  - Recordkeeping:
    - Pittsburgh, RKMS, NAA, VERS, EAD …
  - Multimedia:
    - MPEG-7, SMPTE …
  - Rights management:
    - <indecs>, MPEG-21 …

Draft categorisation (1)



UKOLN

Draft categorisation (2)

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

Implementation

Focus on implementation issues is increasingly important:

- We need to prove the practical value of metadata frameworks and 'outline specifications'
- It can be difficult for implementers to use these as a guide to the design of *real* systems?
- We need to move from the conceptual to the practical, need to move beyond proof-of-concept
- Positive signs:
  - METS/NISO Z39.87
  - OCLC/RLG PREMIS WG looking at implementation strategies for preservation metadata

Creation and capture

Metadata creation/capture:

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

Sustainability (1)

Balance risks with costs:

- There is a perception that metadata creation and maintenance will be expensive
- But costs associated with data recovery are not trivial
- Need to balance the risks of data loss with the cost of creating metadata
  - Robust selection criteria
  - Co-operation between repositories
  - Re-use of existing metadata

Sustainability (2)

Avoid imposing unnecessary costs:

- Avoid large schemas
- Need to identify the *right* metadata ('core metadata'?)

Who pays?

- A more generic issue …

Interoperability (1)

Interoperability is important, e.g.:

- To support the reuse of existing metadata, e.g., on Ingest
- To support the exchange of digital objects between repositories
- "… there is a critical need to develop tools that automatically supply core metadata, extract metadata from resources at ingest, and restructure and manage metadata over time" - (Hedstrom, 2003)

Interoperability (2)

Some problems:

- The need to cope with a wide (and growing) range of metadata standards, object types, formats, etc.
- Heterogeneity
- No prospect of a single standard
- *Practical* interoperability not within easy scope of the OAIS model

Registries (1)

A potential role for registries?

- Format Registries
  - There is "… a pressing need to establish reliable, sustained repositories of file format specifications, documentation, and related software" (Lawrence, *et al*., 2000)
  - DSpace 'bitstream format registry'
  - Typed Object Model (TOM) project
  - IFLA Conference paper (Abrams & Seaman, 2003)

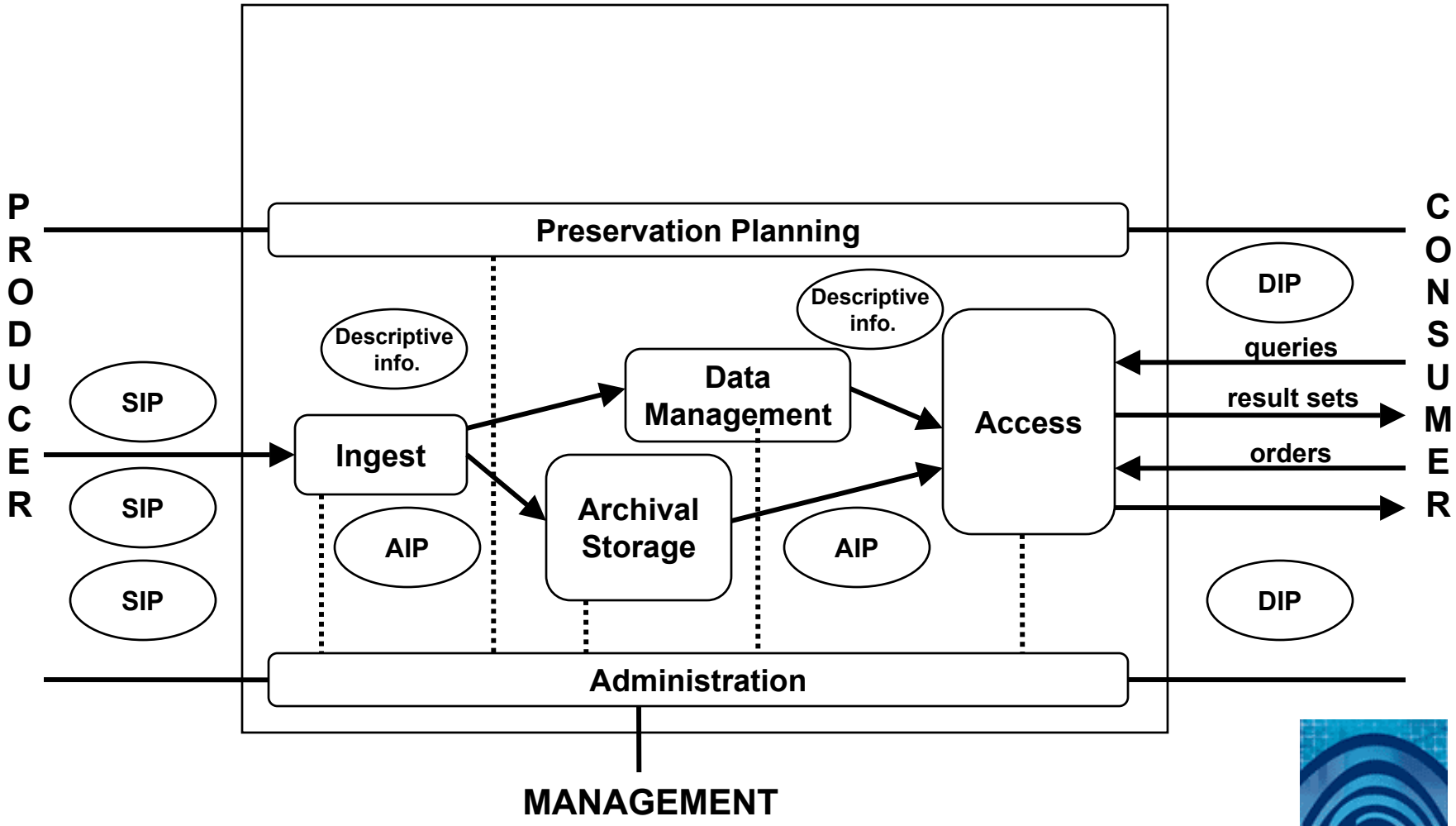
Registries (2)

- Metadata registries
  - "… formal systems that can disclose authoritative information about the semantics and structure of the data elements that are included within a particular metadata scheme" (Heery, *et al*., 2000)
  - Existing registries include the XML.org Registry and Repository (OASIS), and metadata registries set up by DCMI and SMPTE
  - There has been some experimentation with RDF registries as part of Semantic Web development

Registries (2)

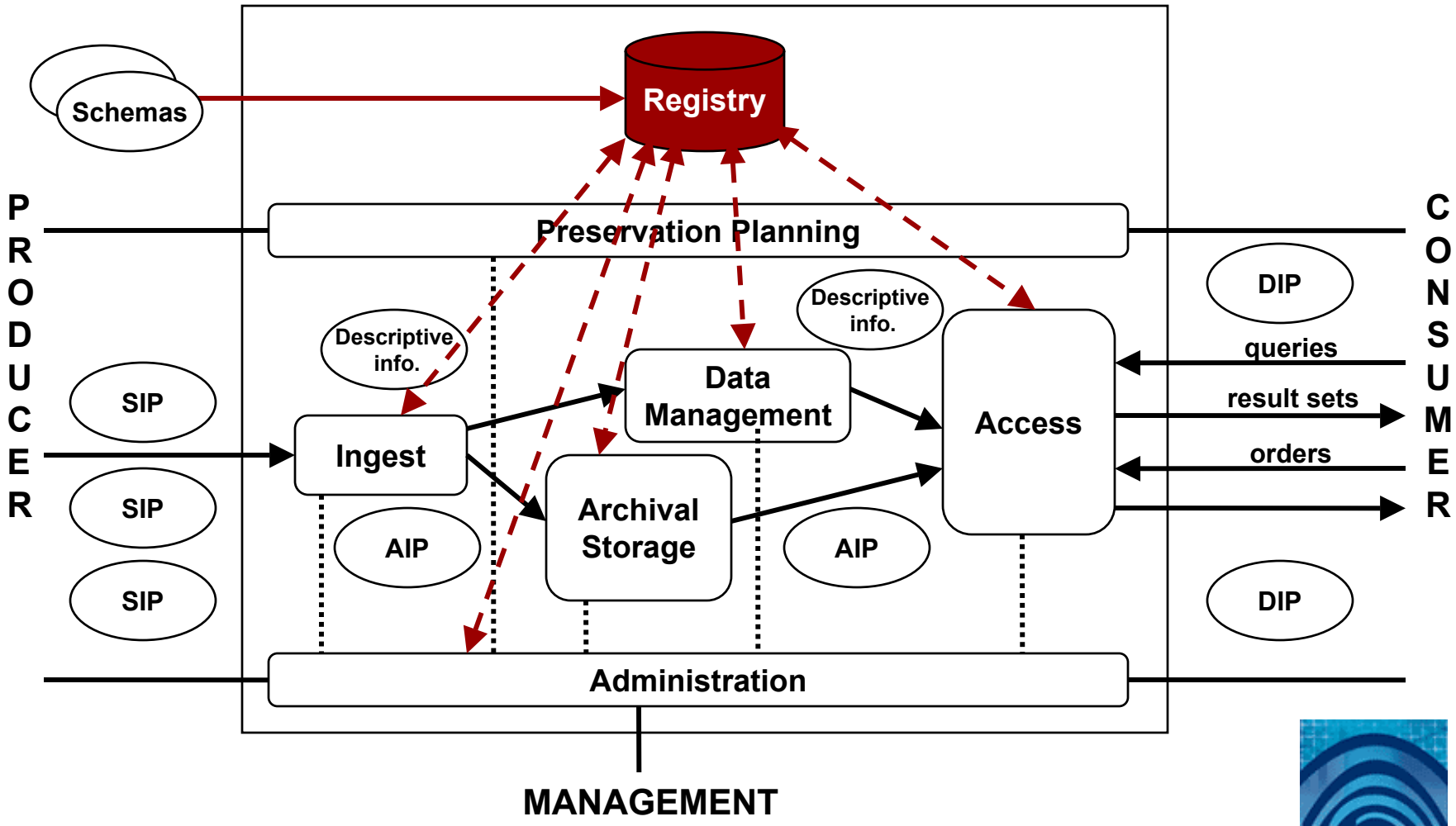
- Registry Functions:
  - Provides support for the ingest process
  - May also provide support for the access function
    - The export of Dissemination Information Packages
    - The exchange of data objects (AIPs?) with other repositories; conversion to exchange standards
  - Can link metadata where there are multiple instances
  - Can help to manage schema evolution

Registry functions



OAIS Functional Entities (Figure 4-1)

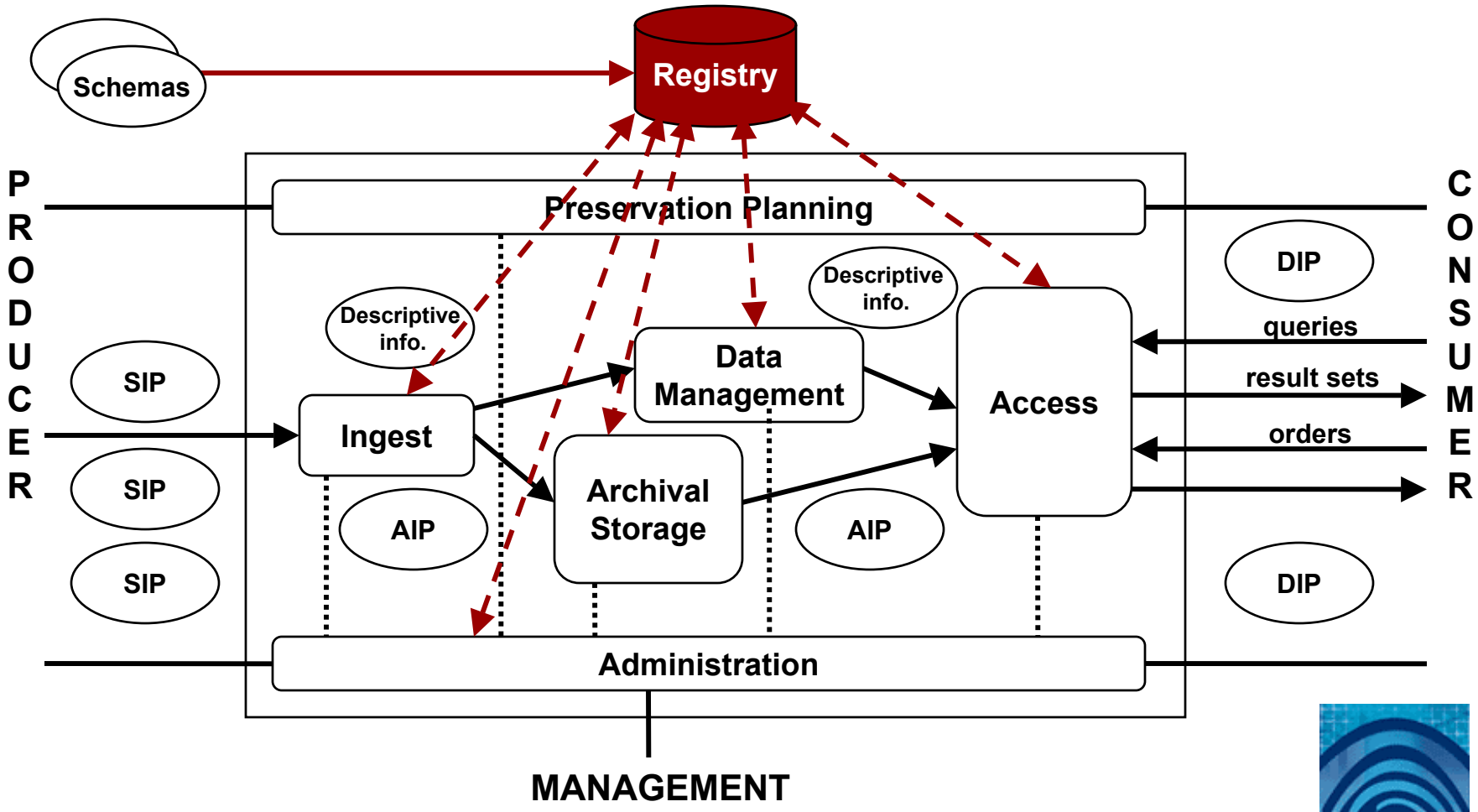
Registry functions



OAIS Functional Entities (Figure 4-1)

UKOLN

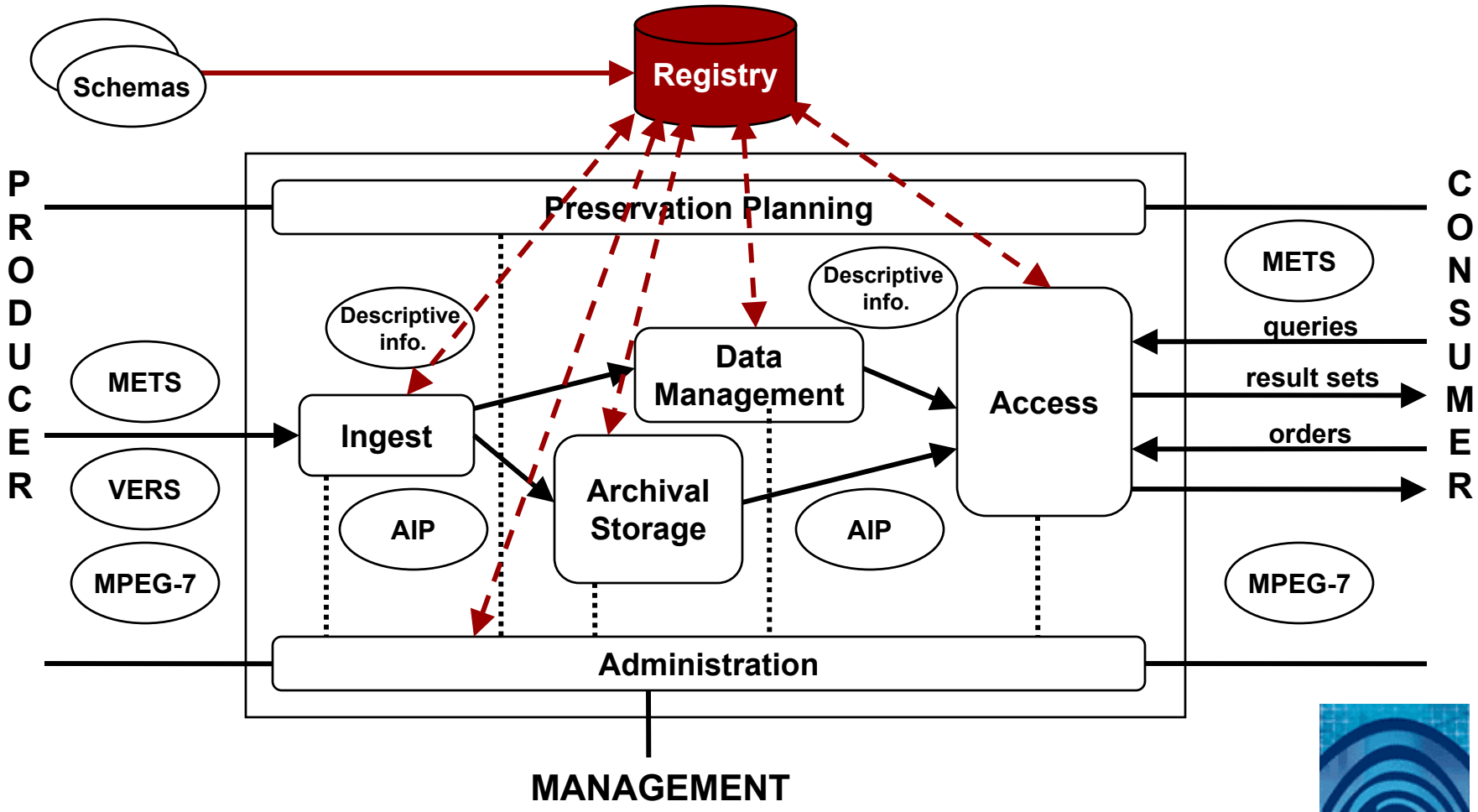
Registry functions



OAIS Functional Entities (Figure 4-1)

UKOLN

Registry functions



OAIS Functional Entities (Figure 4-1)

Open issues (1)

Organisation of registries:

- Registries are part of infrastructure
- Distributed vs. centralised approaches:
  - Hedstrom (2003) suggests that format and metadata registries could/should be 'shared services'
  - Experimental distributed registries are based on Resource Description Framework (RDF)
    - CORES Registry
    - Encourage *re-use* of metadata (the 'application profile' concept)
    - Are other technologies more suitable?
- Who should be responsible?

Open issues (2)

Metadata quality:

- Standards often deal with metadata semantics, but not always with 'content rules'
- Recent experience with use of unqualified Dublin Core by OAI data providers suggests that metadata quality varies widely, e.g.:
  - DC underutilized, e.g. 5 of 15 elements used 71% of the time, many records have just 'title' and 'creator' elements (Ward, 2003)
  - Some quality problems with records being imported before their refinement by libraries (Halbert, 2003)
- Authority control, de-duplication, etc.

Open issues (3)

Different data models:

- How does the OAIS model fit into other data models being developed for (digital) objects?
  - Examples:
    - Functional Requirements for Bibliographic Records (FRBR) - IFLA
    - ABC Ontology and Model - Harmony project
    - CIDOC Conceptual Reference Model (CRM)
    - …

Summing up

- Implementation issues:
  - A need to focus on the practical issues of implementing preservation metadata standards within *real* systems
  - Then feed what is learnt through this back into the schema design (iterative process)
  - If it doesn't work, start again …
- Interoperability:
  - For reuse and exchange of metadata
  - Possible role for format and metadata registries - but the concept needs extensive testing (and registries are not a panacea)

Acknowledgements

UKOLN is funded by Resource: the Council for Museums, Archives and Libraries, the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath, where it is based.

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