OAIS & File Format Repositories

Robert Sharpe
Tessella Support Services

24 August 2004

ICA 2004, Wien

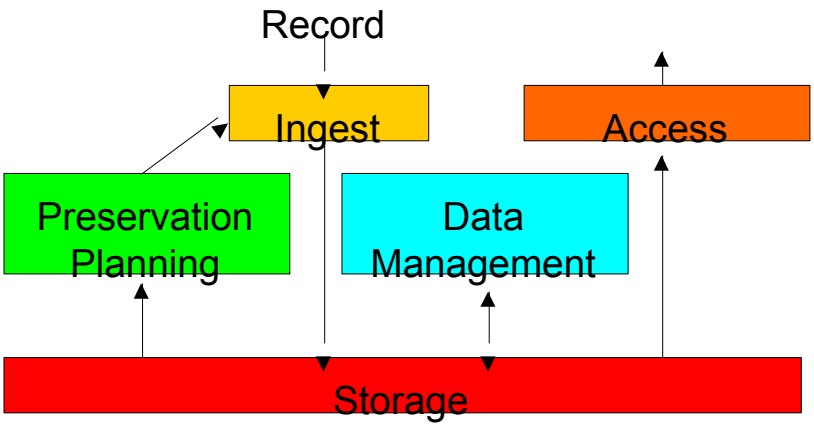
Contents

- Building an OAIS-compliant archive:
  - Stores digital files.
  - Plus associated metadata.
- Building a file format repository:
  - Information on each format in archive.
- Interaction between the two:
  - Enabling preservation planning.

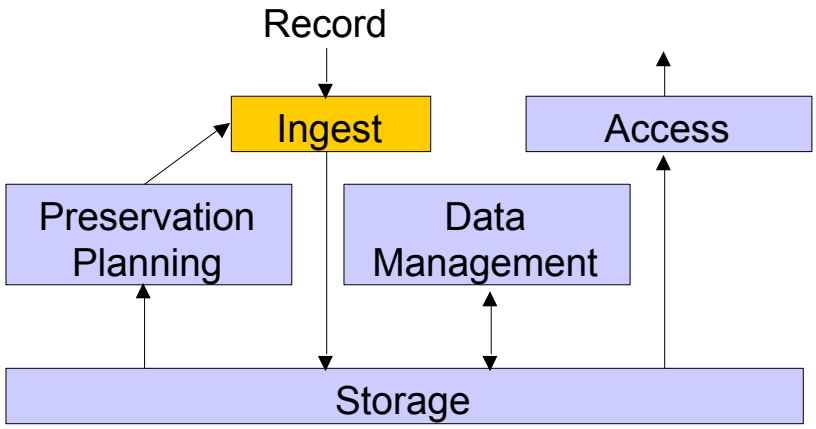
OAIS framework

- Framework come out of space community (NASA et al.).
- Now an ISO standard.
- Useful split of problem:
  - Ingest.
  - Data management.
  - Storage.
  - Access.
  - Preservation planning.
  - Administration.
- Used as a blueprint for real archives.

OAIS framework



Ingest



Ingest Tasks

- Selection.
- Determine record structure:
  - Logical.
  - Physical.
- Set metadata:
  - Finding aids.
  - Technical / preservation.
- Verify, load & move to storage.

Ingest: Selection

- Need close relationships with record suppliers.
- Ideal: From ERMS with metadata with files in set format
- Reality: Take what given by series.
- Easier if can select by format?:
  - Often not possible.

Ingest: Record structure example

- e.g., archive minutes of a committee
- Logical Structure:

Important Committee
- Meetings
  - Meeting 1
  - Meeting 2
  - Meeting 3
  - …

Ingest: Record structure example

- Scenario A: 1 Word doc per meeting.
- Physical & logical structure coincide.

Important Committee
- Meetings
  - Meeting 1 → Word 1.DOC
  - Meeting 2 → Word 2.DOC
  - Meeting 3 → Word 3.DOC
  - …

Ingest: Record structure example

- Scenario B: Keep in a database.

Important Committee
- Meetings → Database.MDB
  - Meeting 1
  - Meeting 2
  - Meeting 3
  - …

Ingest: Record structure example

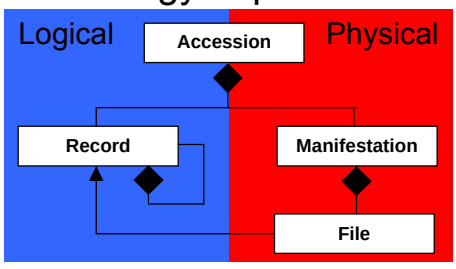
- Scenario C: Part of a committee-wide, complex system

Important Committee → Lots of files

- Meetings
  - Meeting 1
  - Meeting 2
  - Meeting 3
  - …

Ingest: Record structure

- Logical structure:
  - Technology independent.
- Physical structure:
  - Technology dependent.



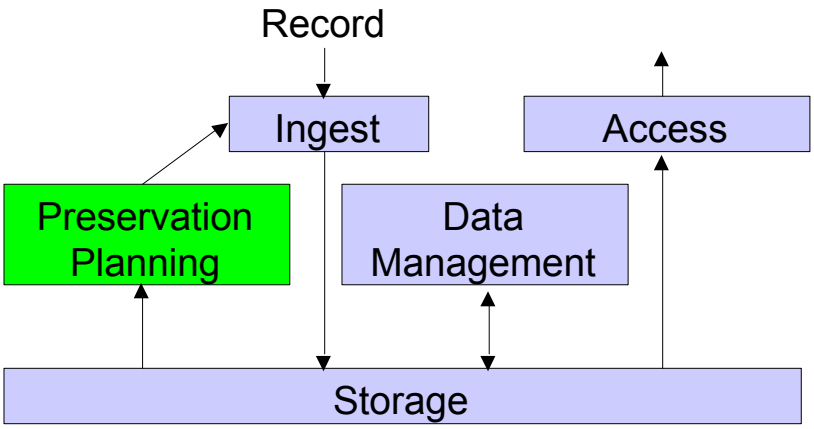
Ingest: Set Metadata

- Set metadata for each entity:
  - Accession:
    - Acquisition information etc.
  - Record:
    - Context, indexing, access conditions etc.
  - Manifestation:
    - Hardware, O/S, Application software needed etc.
  - File:
    - Size, format etc.

Ingest: Metadata extraction

- File metadata relatively easy:
  - 3rd party software detects most formats.
  - Issue of validity / corruption?
- Manifestation metadata:
  - Ideally look up based on file formats.
  - Need file format repository.
- Record/accession metadata :
  - Ideally, take rest from ERMS.
  - Otherwise slow “archaeology”.
  - Or need automatic extraction tools.

Preservation Planning



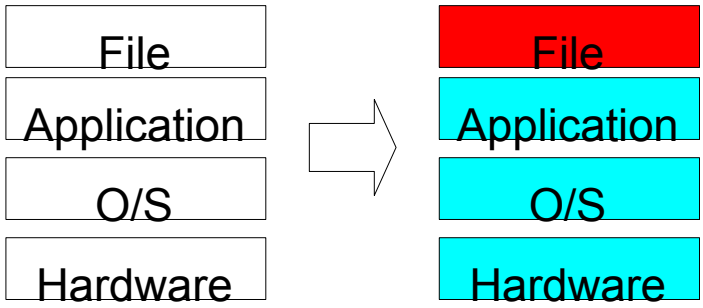
Preservation Planning

- Move to new technology.
- Aim is to retain:
  - Context.
  - Content.
  - Structure (logical / physical).
  - Appearance (look & feel).
  - Behaviour (macros, programs).

Preservation Planning

- Two realistic options:
  - Migration
  - Emulation
- Need file format information to determine best method:
  - File format repository

Preservation Planning: Migration



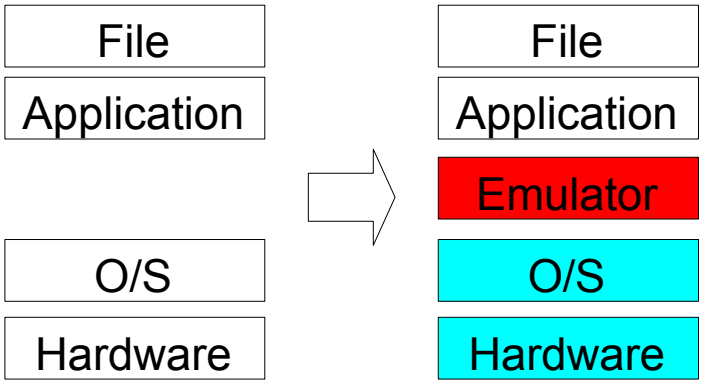
Preservation Planning: Migration

- Context: preserved (metadata)
- Content: should be unaltered.
- Structure:
  - Logical records preserved.
  - Physical structure may change.
- Appearance: harder, some loss.
- Behaviour: unlikely to be preserved.

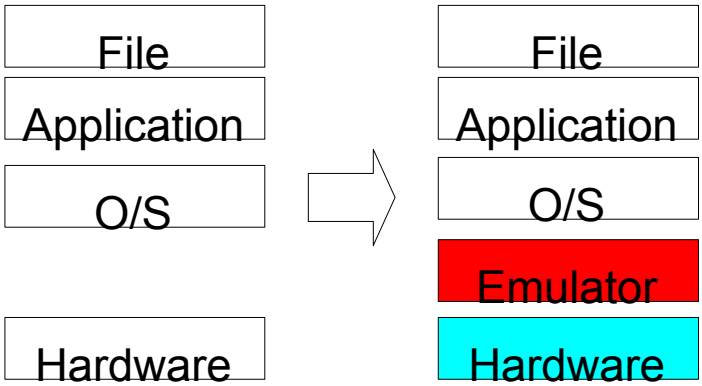
Preservation Planning: Migration

- Transformation engines exist:
  - Convert popular formats to XML.
  - Can convert XML to HTML.
- Are they up to archival standards?
  - Maybe not but better than nothing?
  - Keep the original so can always improve.
  - If there is a need, the software will get better!
  - Does it matter if they are open source?

Preservation Planning: O/S Emulation



Preservation Planning: Hardware Emulation



Preservation Planning: Emulation

- Should retain:
  - Appearance.
  - Behaviour.
- But:
  - Difficult to write generic emulator.
  - Not yet proven.

File format repository

- Want information on formats.
- Need to hold information on:
  - Software (hierarchy of components).
  - Manufacturers.
  - Hardware.
  - Encoding methods.
  - Compression algorithms.

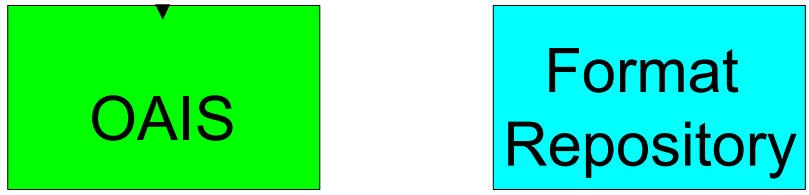
File format repository

- Need accurate information on support dates:
  - Determine when need to do something.
- Hold proposed action paths:
  - E.g., Migrate using a given transformation engine to new format.
- Need to tell OAIS repository!

System interaction

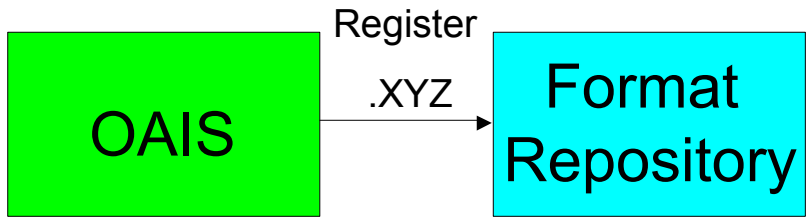
- OAIS repository receives files

New File.XYZ



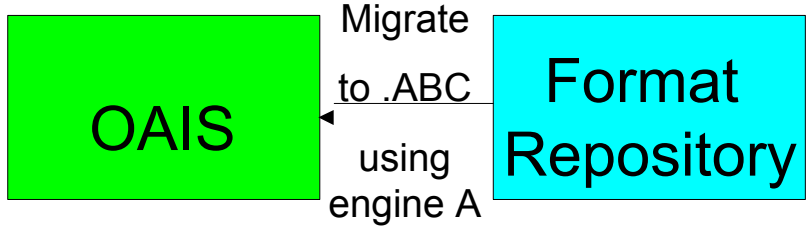
System interaction

- New format so registers interest



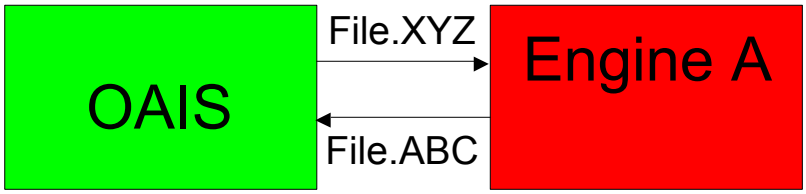
System interaction

- Repository sends back proposed actions (time delay)



System interaction

- OAIS system schedules actions
- Performed & saves new versions.



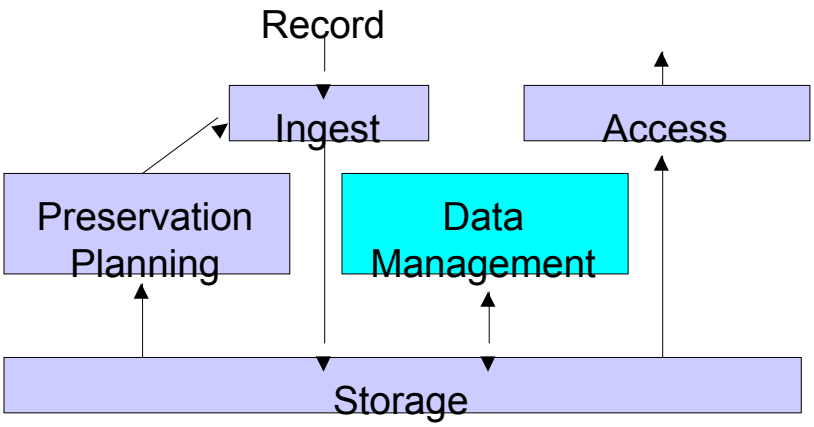
Conclusions

- Can build OAIS repositories.
  - Been done (e.g., UK Digital Archive).
  - More planned.
- Can build file format repositories.
  - Been done (e.g. PRONOM).
  - More planned.
- Can get them to interact.
  - Plans.
- Can automate digital preservation.
- It is happening!

Acknowledgments

- UK National Archives:
  - Digital Archive – visit Kew:
    http://www.nationalarchives.gov.uk/preservation/digitalarchive/
    - Pilgrim Trust award winner.
  - PRONOM:
    http://www.nationalarchives.gov.uk/pronom/
- NARA's ERA program:
  http://www.archives.gov/electronic_records_archives/

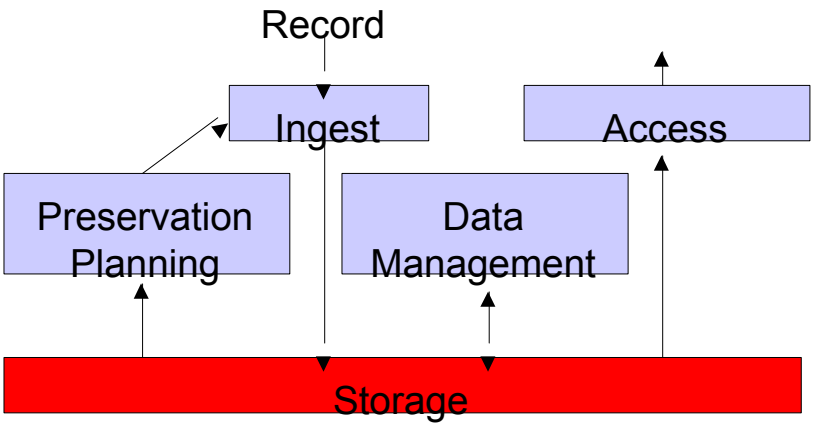
Data Management



Data management

- Allowed to edit metadata:
  - All actions audited.
- Can NOT edit file contents.
- Can add :
  - Accruals
  - Redacted versions.
  - Migrated files.

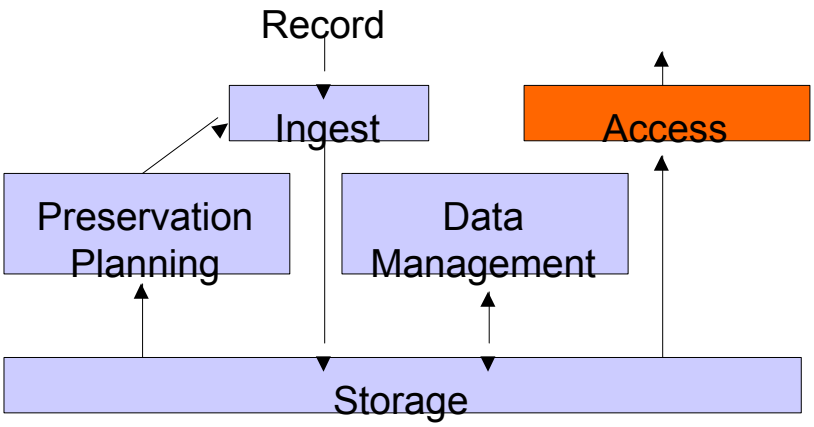
Storage



Storage

- Store:
  - Metadata.
  - Digital files.
  - Could encapsulate, e.g. VERS?
- Managed environment.
  - Continuous program of checks.
  - Controlled media migration.
  - Backup & restore.

Access



Access

- Search:
  - Search by phrase:
    - Options to restrict date, dept.
  - Browse.
  - Can view metadata.
- Dissemination:
  - Choose logical record to download.
  - Get multiple files (ZIP).
- Demonstration.