

Active Preservation

A practical approach to long term
digital preservation

James Carr

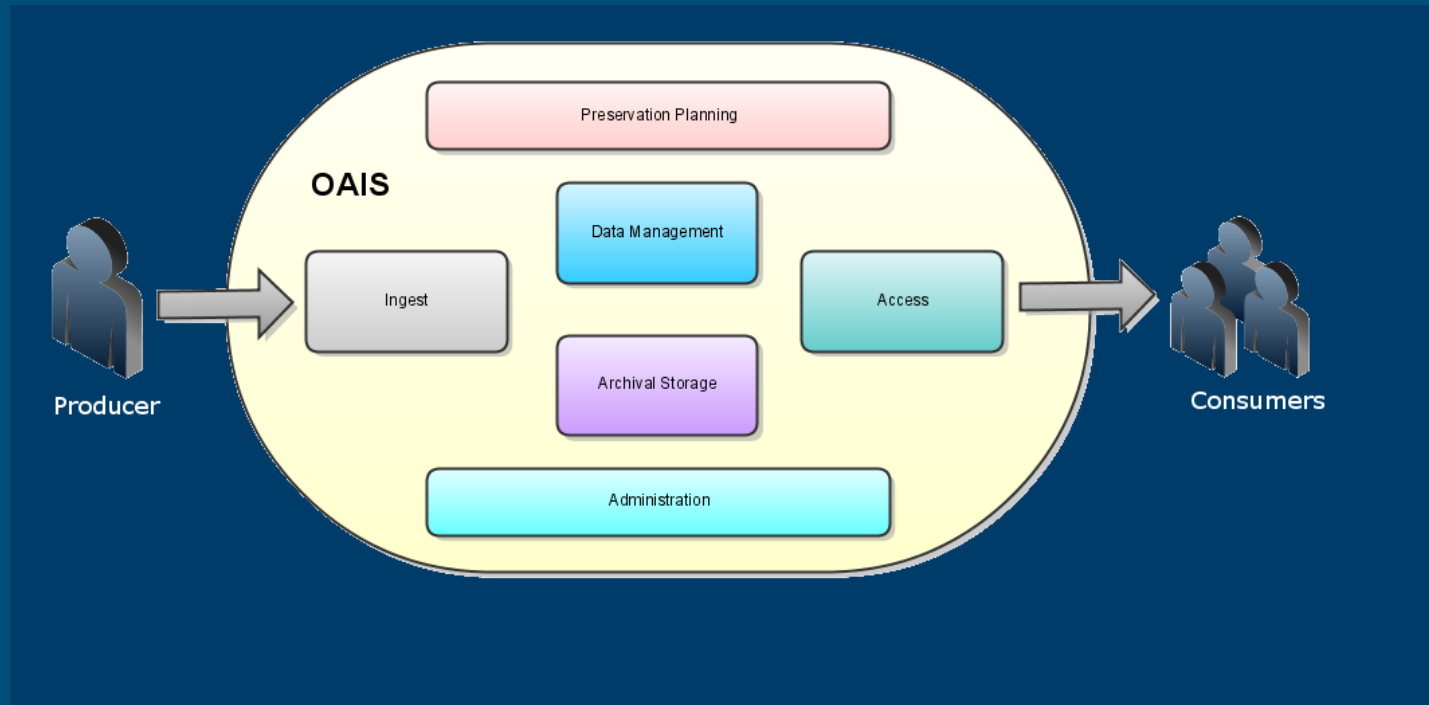
March 23rd 2010

Contents

- What is “Active Preservation”?
 - History
- Nature of digital records:
 - Why is long-term preservation hard?
- Active Preservation:
 - Technical Registry
 - Characterisation
 - Preservation Planning
 - Migration
- The future

What is “Active Preservation”?

- OAIS



- Active Preservation = Preservation Planning (automated)
- BUT influences all functional entities

The Nature of Digital Records

Characteristic

- Context
- Content
- Appearance
- Behaviour
- Physical Structure
- Conceptual Structure

Paper

Metadata
Paper
Paper
None
Arrangement
Arrangement

Electronic

Metadata
File format
Format/SW/HW
Format/SW/HW
File system hierarchy
Format/SW/HW

Digital Preservation – Well known issues

- Can not read digital records directly:
 - Rely on file formats
 - Rely on application software
 - Rely on operating system
 - Rely on hardware.
 - Obsolescent within information object lifetime
- Preservation strategy must rely not just on preserving the original but also “lossless” transformation:
 - Migration
 - Emulation
- Active Preservation deals with this

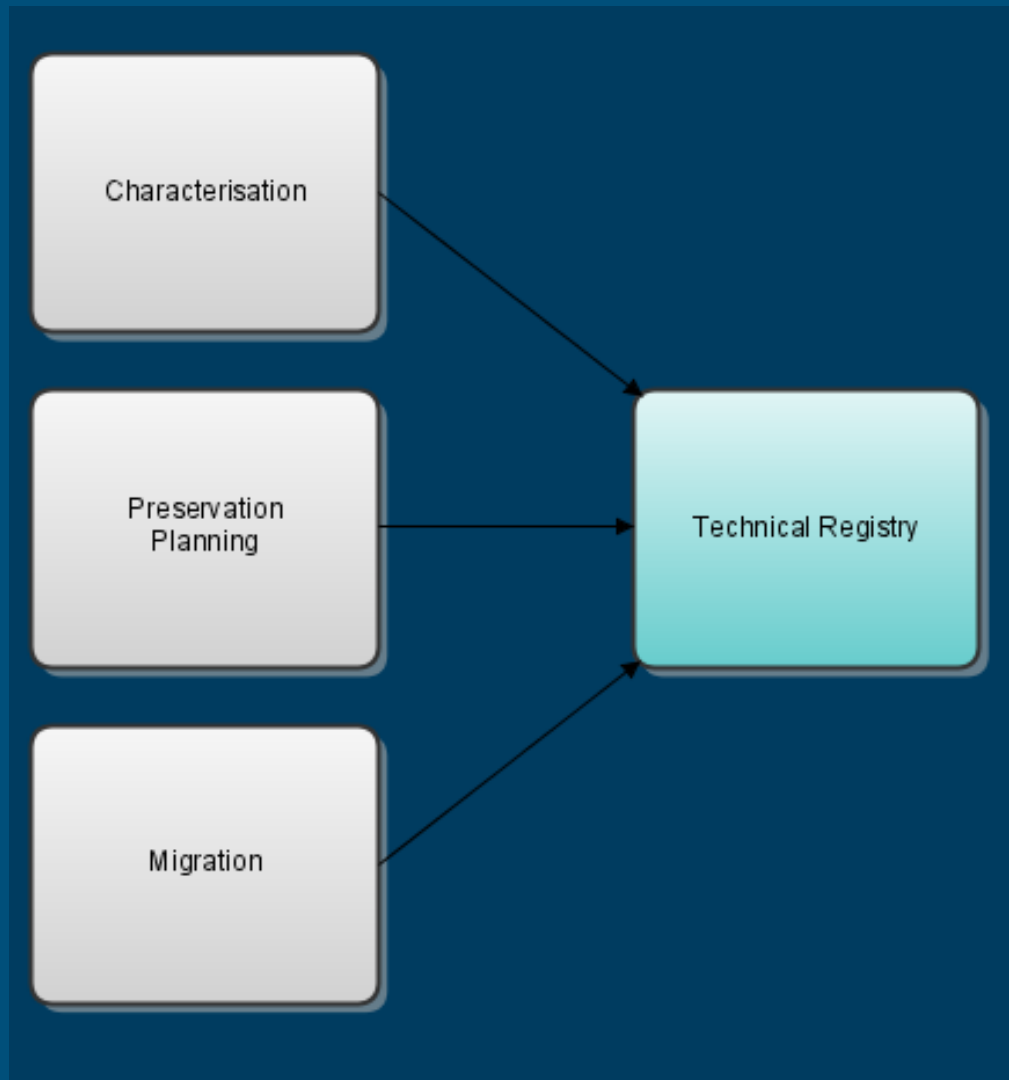
Digital Preservation – Less well known issue

- Break link between physical and conceptual structures:
 - e.g., Web site
- Really two structures:
 - Conceptual (digital records or information objects):
 - Understood by humans
 - Technology independent
 - Needs to be preserved
 - Physical (digital objects):
 - Understand by machines
 - Technology dependent
 - May need to be migrated / emulated
- Active Preservation deals with this too

Active Preservation - Overview

- Step 1: Characterisation:
 - What have I got physically?
 - What have I got conceptually?
- Step 2: Preservation planning
 - Decide what is at risk?
 - What should I do about it?
- Step 3: Preservation Action
 - Perform plan
 - Include re-characterisation to validate
 - Currently just migration but will include emulation

Active Preservation: Technical Registry



Technical Registry - Overview

- All information needed to decide how active preservation is to be performed:
 - Factual information (formats, software, properties etc.)
 - Also policy (risk criteria, preferred migration pathways etc.)
- Exists with various names:
 - PRONOM (UK National Archives):
 - <http://www.nationalarchives.gov.uk/PRONOM>
 - Planets Core Registry (PCR)
 - Planned to be the seed of The Unified Digital Formats Registry (UDFR)

Technical Registry

- Factual Information

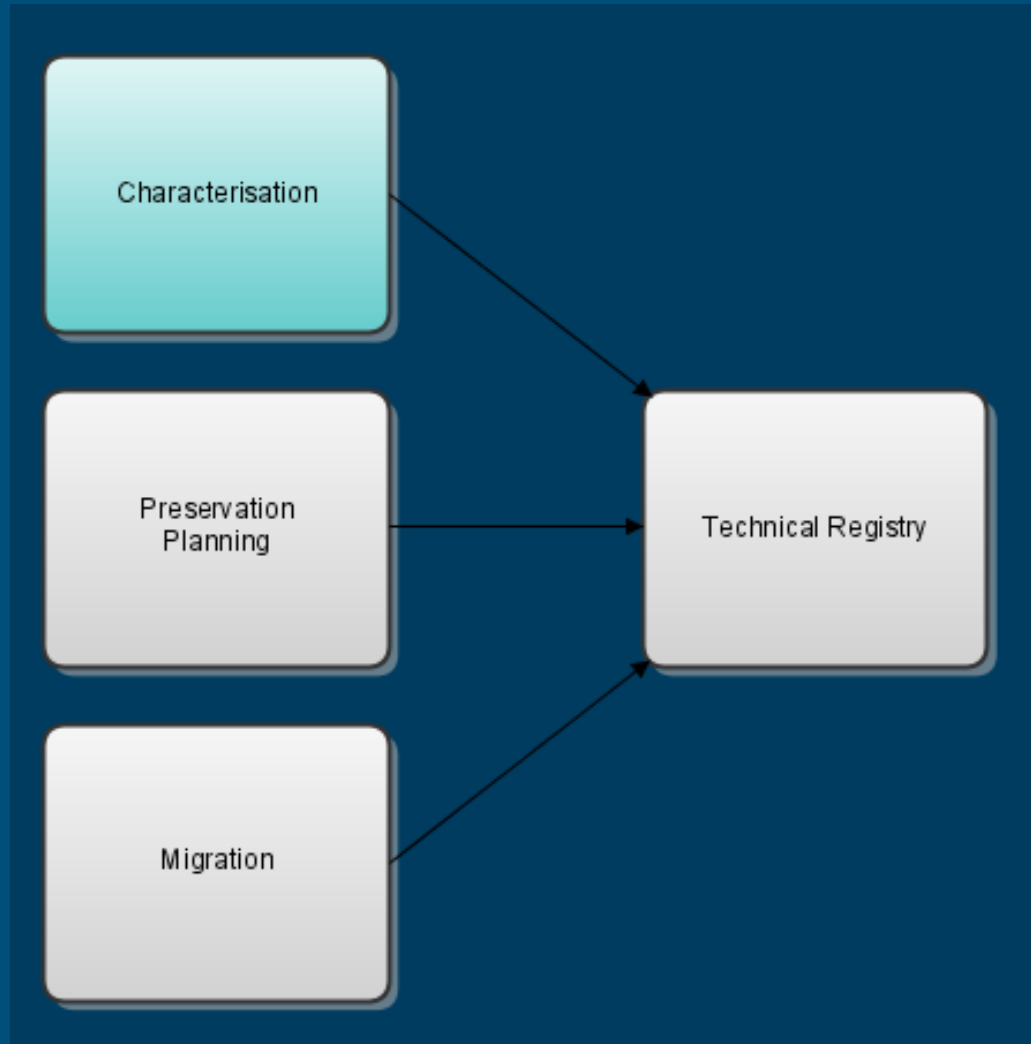
- File formats
- Software applications
- Hardware
- Software capabilities (e.g., create format / render format)
- Software dependencies
- File Properties
- Software Tools
- Migration pathways

Technical Registry

- Policy Information

- Decide which properties to measure?
- What tools to use?:
 - Identification, validation, property extraction
- Associate risks with:
 - Formats (Format Inherent Risk)
 - Format properties (Format Instance Risk)
- Which component properties are invariant?
 - Degree of tolerance?
- Which migration pathway to follow under which circumstances?

Active Preservation: Characterisation



Characterisation - Overview

- What have I got physically (file level)?
 - Format properties that might indicate obsolescence etc.
 - Technology dependent properties

- What have I got conceptually (information object level)?
 - Conceptual Structures (known as components)
 - Capture the “Essential characteristics” of a record
 - Technology independent properties

File Characterisation

- Identification:
 - Single tool: e.g., DROID
- Validation:
 - Tool depends on format: e.g, Jhove for PDF
- Property extraction:
 - Tool depends on format, e.g, apache POI for Office formats
 - Properties to measure depends on format/tool combination
- Detect embedded objects:
 - Tool depends on format, e.g, ZIP
- All automatic, controlled by policy in the Registry

Conceptual Characterisation

- Conceptual vs. physical structures
- Receive physical:
 - In today's technology
- Identify conceptual entities (components):
 - Including links / dependencies
 - Measure "essential characteristics"
- Retain conceptual structure:
 - Accept change in physical structure

Conceptual Characterisation - Example

- Receive physical:
 - Home.html
 - Style.css
 - Logo.gif
 - Page1/Page1.html
 - Page1/Image.jpg
 - Page2/Page2.html
 - Page2/Image.png
 - Page3/Page3.html
 - Page3/Document.pdf

Conceptual Characterisation - Example

- Identify components:

Physical

Page2/Page2.html

Uses Style.css

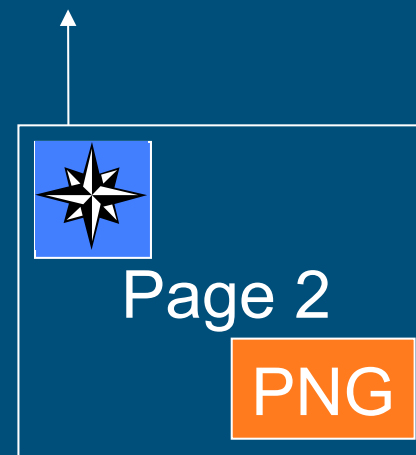
Embeds Logo.gif

Embeds Page2/Image.png

+ Link to Home

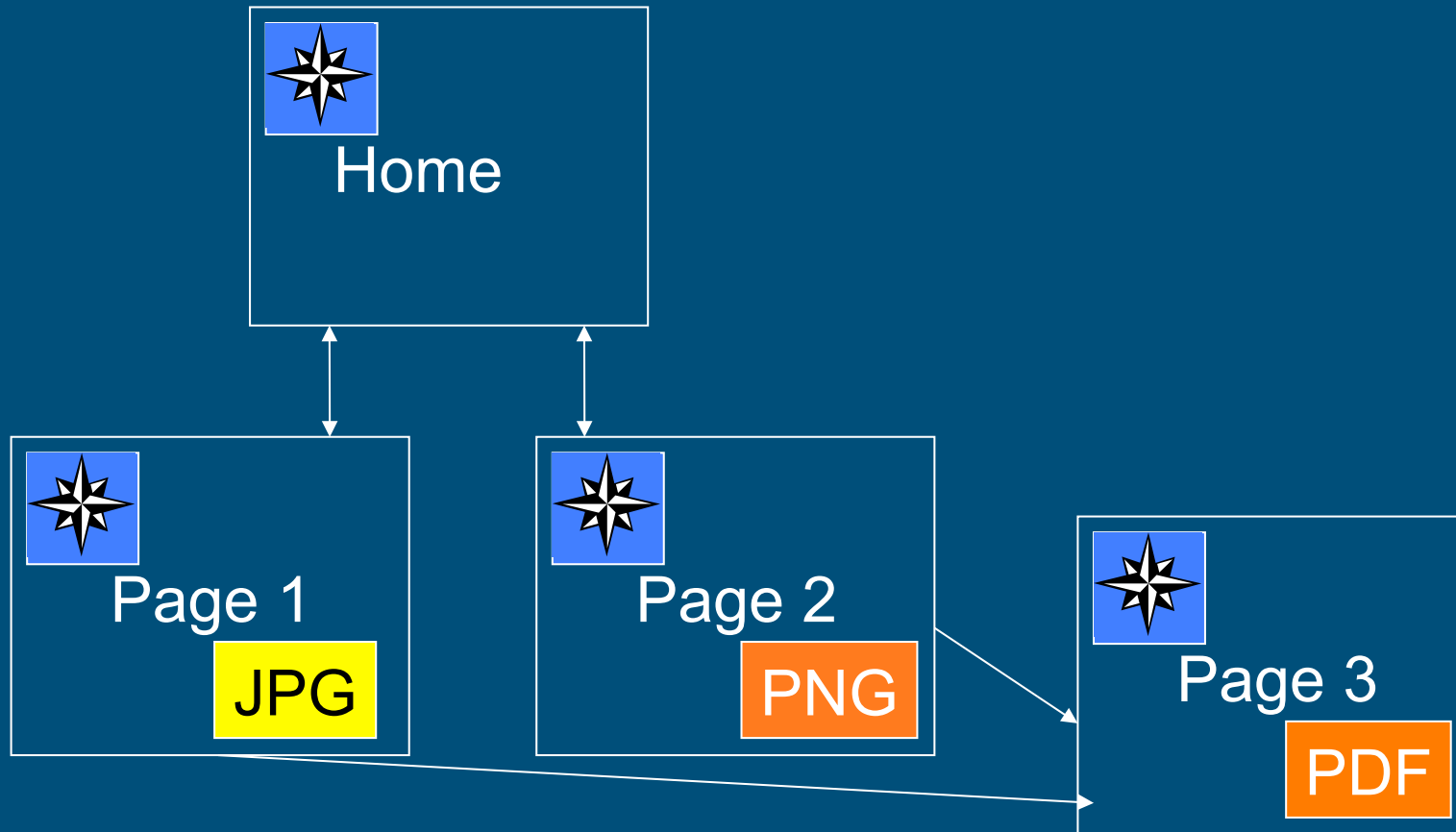
+ Link to Page 3

Conceptual



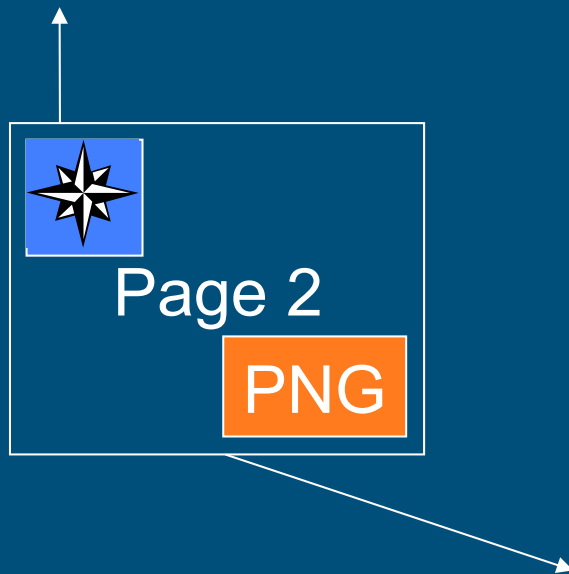
Conceptual Characterisation - Example

- Work out conceptual structure:



Conceptual Characterisation - Example

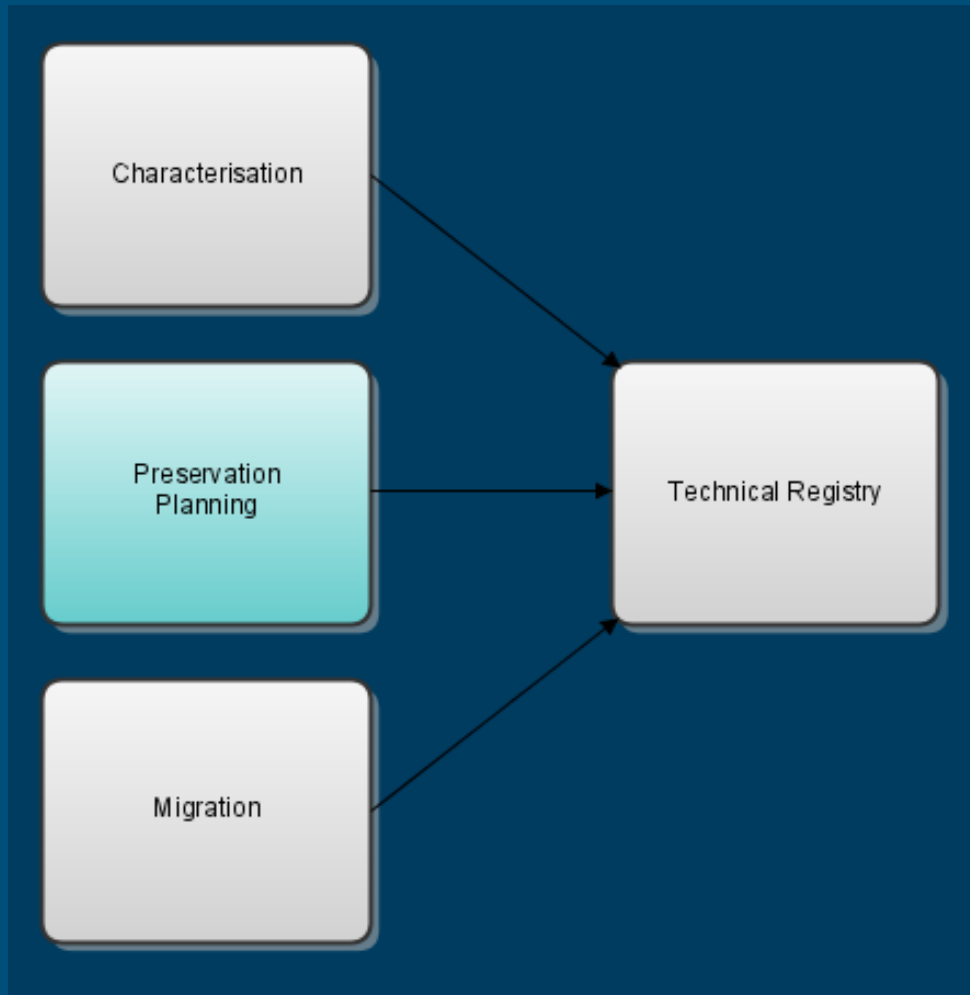
- Identify characteristics:



- Title
- Link to home page
- Link to Page 3
- Contains image:
 - Height
 - Width
- Contains image:
 - Height
 - Width

- Conceptual characteristics **not** linked to technology

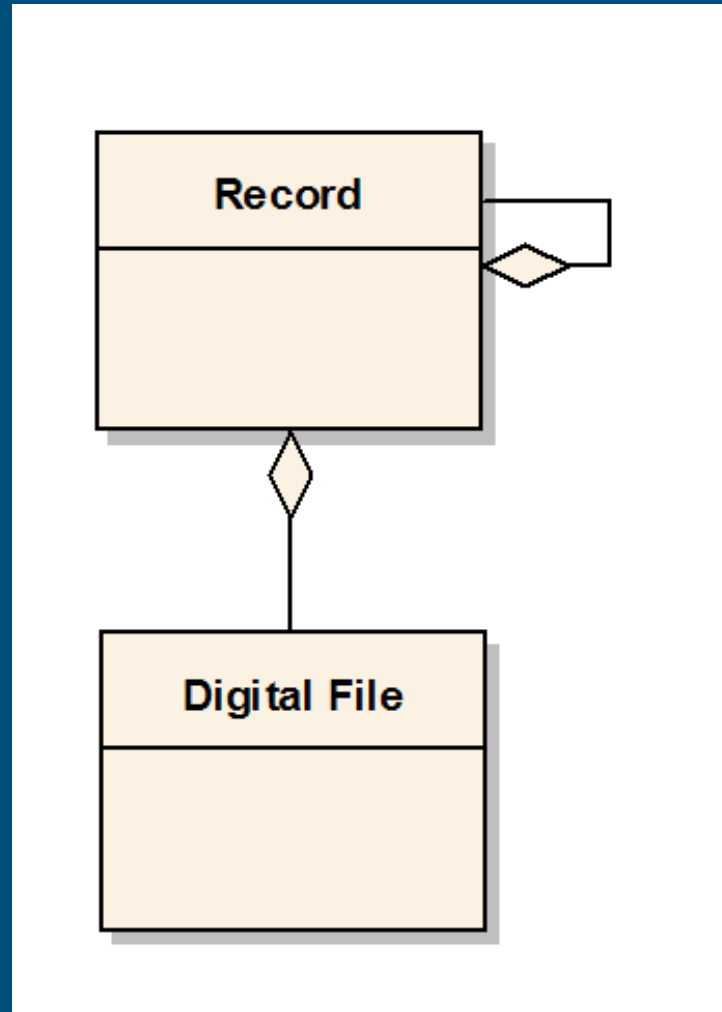
Preservation Planning



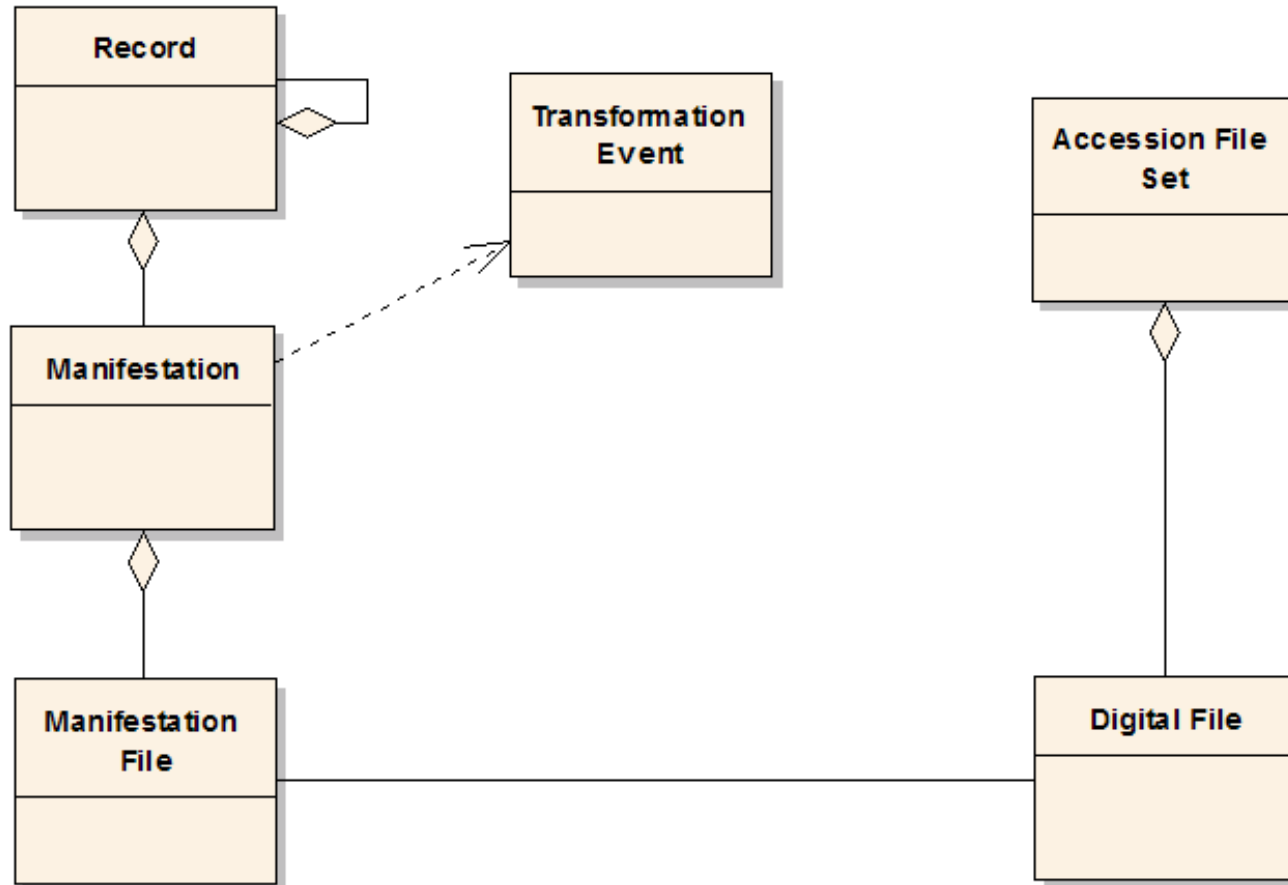
Preservation Planning

- Know format / property combinations at risk
- Decide if files at risk:
 - Base off format and properties
- Know which “active” manifestations are associated with files
- Hence, decide which records need attention

Digression into Data Modelling



Adding manifestations



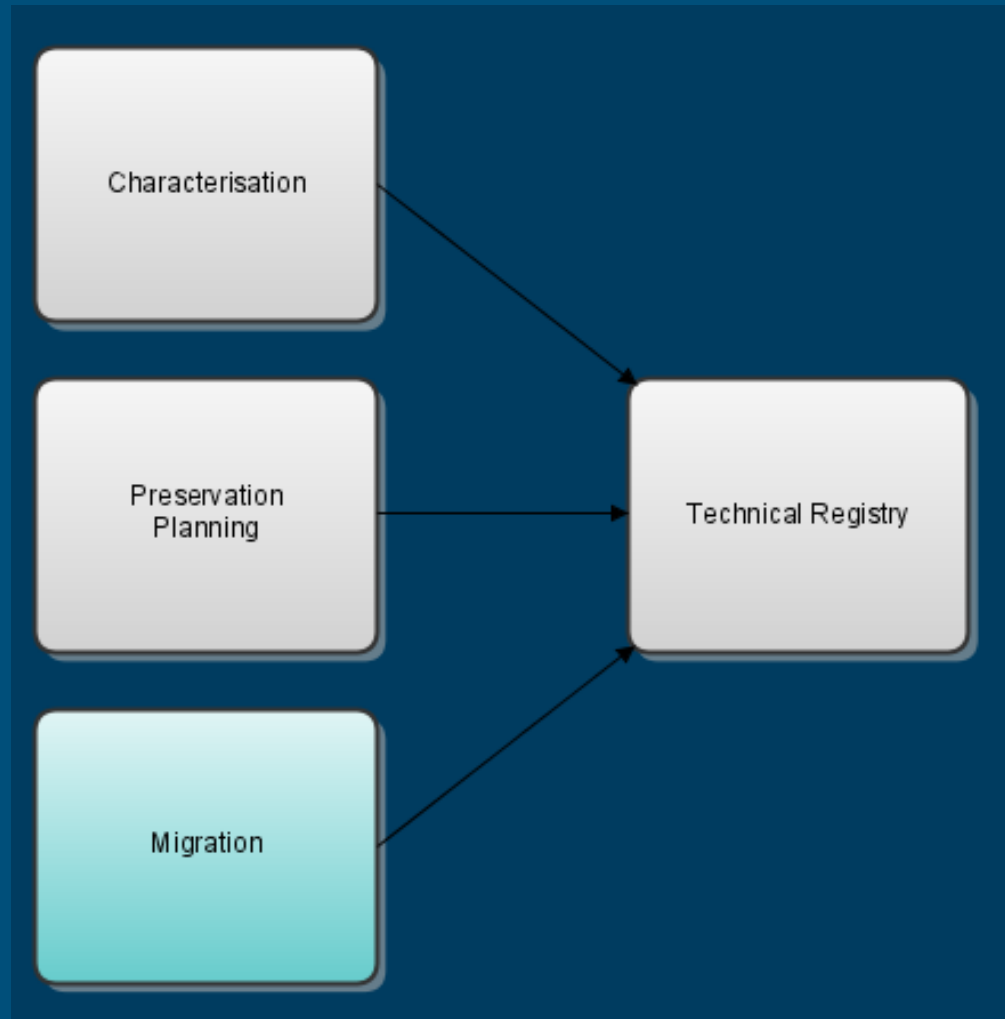
Preservation Planning

- Which new formats can we migrate to?
- Will this reduce the risk?
- Are we creating preservation or presentation manifestations.
- Which tools are available?

Preservation planning output

- XML document describing the records, components and files at risk and the proposed migration pathways for each format.

Migration



Migration

- Execute the Preservation plan
 - Conceptual component is atomic level of migration:
 - Consume a set of files
 - Create a set of files
 - Not necessarily 1-to-1.
 - New files are created, but the same components remain before and after migration.
- Verification:
 - Characterise new files
 - Re-characterise the new manifestation of the information object
 - Check component structure not changed
 - Compare essential characteristics before and after
 - Can also run specific tool, e.g., image comparison tool
- If passes, ingest new manifestation

Active Preservation: Summary

- Current Framework
 - Have tools for common formats
- Future:
 - Develop best practice
 - Wrap more characterisation and migration tools: Planets, Xena, NLNZ etc.
- Questions?