Active Preservation

Sun PA SIG, Malta 24th - 26th June 2009

Tessella
Technology & Consulting

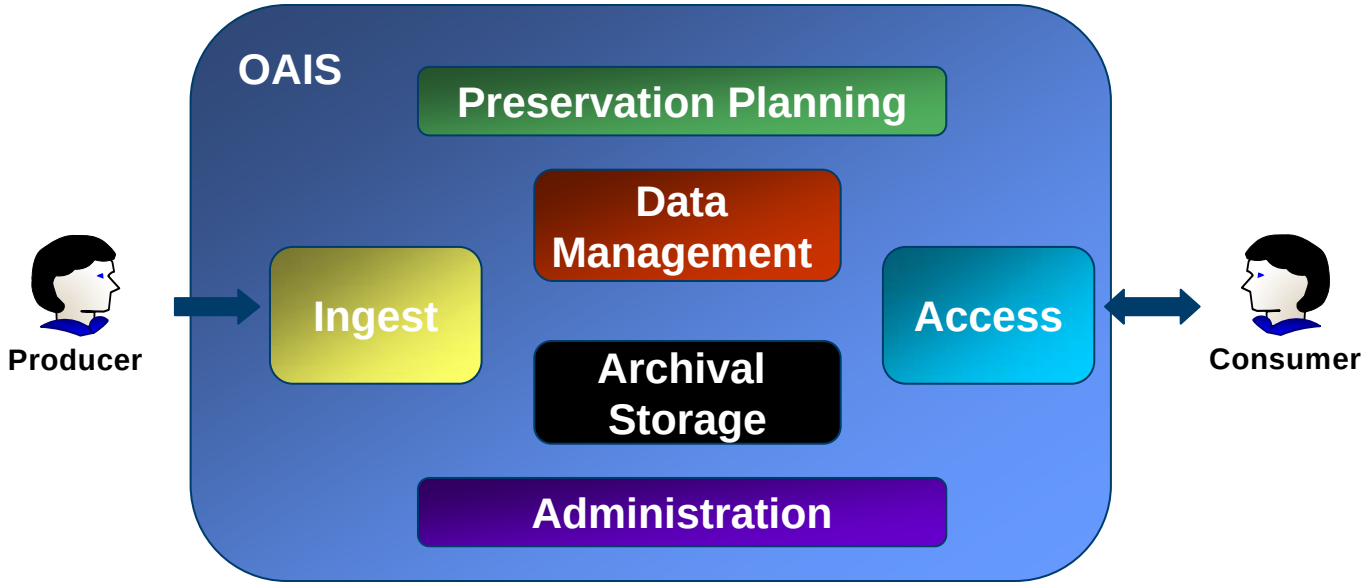
Contents

- 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

Tessella
Technology & Consulting

What is “Active Preservation”?

- OAIS



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

Tessella
Technology & Consulting

History

- Tessella worked with UK National Archives:
  - PRONOM (2001+)
  - Digital Archive (2002-3)
  - DROID (2004)
  - Seamless Flow (2006+)
- Award-winning!
- Developed into Safety Deposit Box (SDB):
  - Full OAIS archiving and preservation solution
  - Incorporates TNA's technology under licence
  - Plus input from EU Planets project
- More customers…

Tessella
Technology & Consulting

World Leading Digital Archive Solutions

The British Library

UK National Archives

Wellcome Trust

Rotterdam City Archive

Dutch National Archives

Swiss Federal Archives

US National Archives and Records Administration

Malaysian Archives

Tessella
Technology & Consulting

Digital Preservation – Well known issue

- Can not read information objects directly :
  - Rely on file formats
  - Rely on application software
  - Rely on O/S
  - 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

Tessella
Technology & Consulting

Digital Preservation – Less well known issue

- Break link between physical and conceptual structures:
  - e.g., Web site
- Really two structures:
  - Conceptual (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

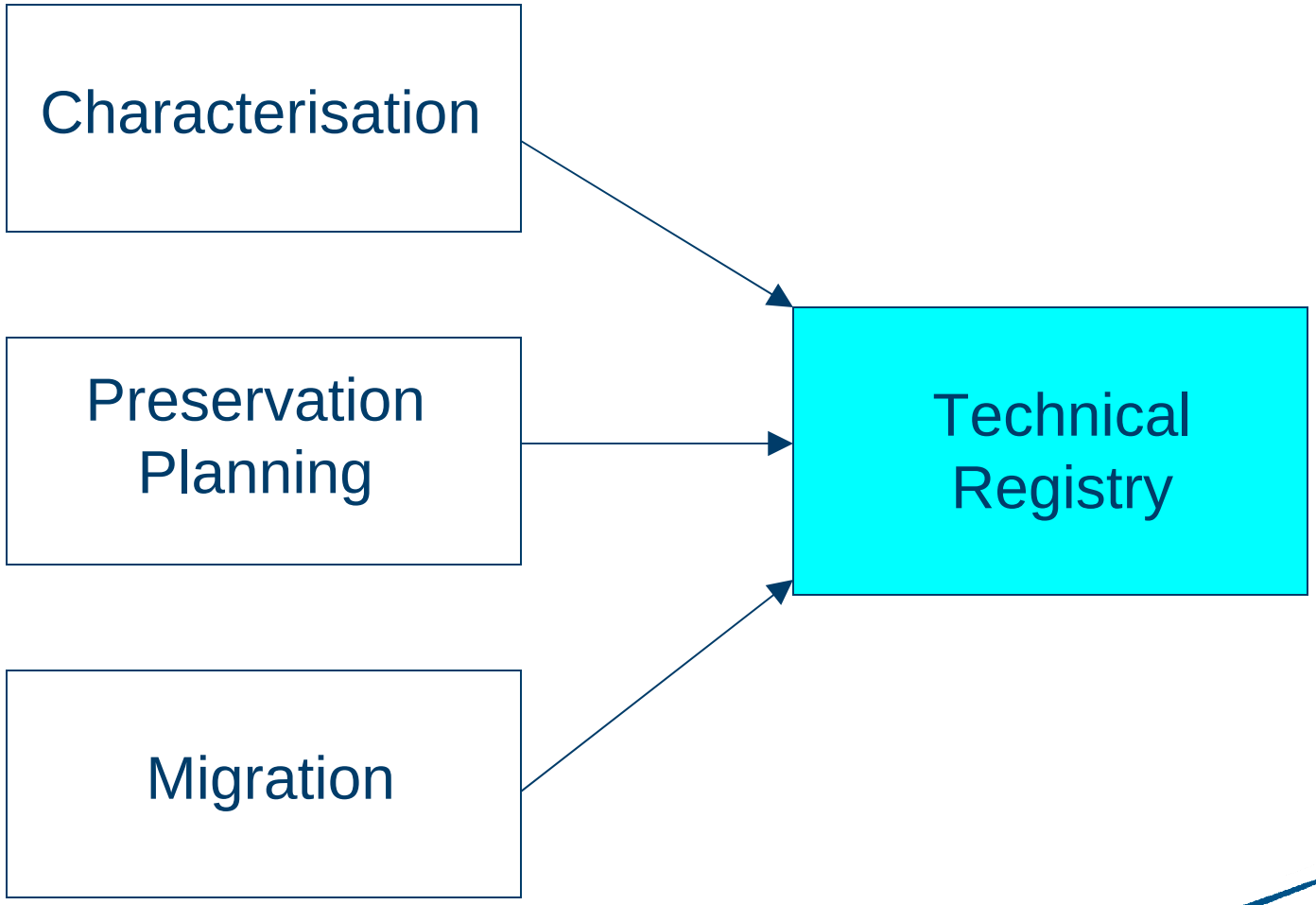
Tessella
Technology & Consulting

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

Tessella
Technology & Consulting

Active Preservation: Technical Registry



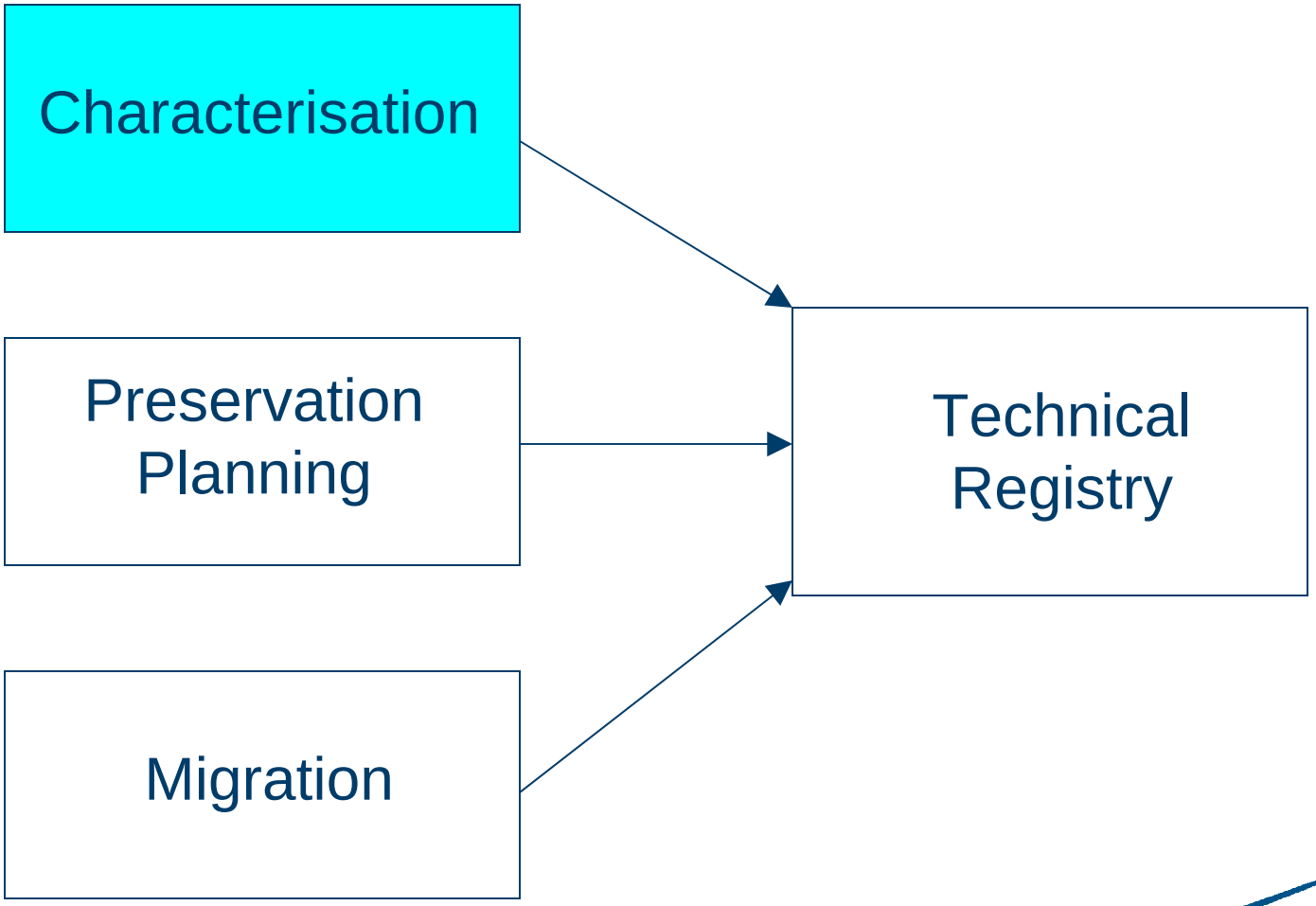
Tessella
Technology & Consulting

Technical Registry - Overview

- All information needed to decide what active preservation to perform:
  - Factual information (formats, software, properties etc.)
  - Also policy (risk criteria, preferred pathways etc.)
- Exists with various names:
  - PRONOM (UK National Archives):
    - www.nationalarchives.gov.uk/PRONOM
  - Planets Core Registry
  - Planned to be the seed of UDFR

Tessella
Technology & Consulting

Active Preservation: Characterisation



Tessella
Technology & Consulting

Characterisation - Overview

- What have I got physically (file level)?
  - Format, properties that might indicate obsolescence etc.
  - Technology dependent
- What have I got conceptually (information object level)?
  - Conceptual Structure
  - “Essential characteristics”
  - Technology independent

Tessella
Technology & Consulting

File Characterisation

- Identification:
  - Single tool: e.g., DROID
- Validation:
  - Tool depends on format: e.g, Jhove
- Property extraction:
  - Tool depends on format, e.g, Jhove
  - 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

Tessella
Technology & Consulting

Conceptual Characterisation

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

Tessella
Technology & Consulting

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

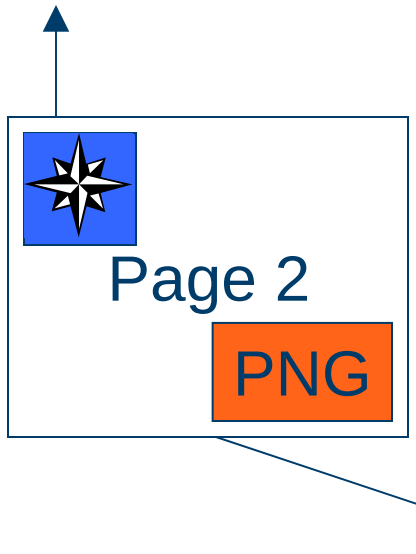
- Identify components:

Physical

Page2/Page2.html

Uses Style.css
Embeds Logo.gif
Embeds Page2/Image.png
+ Link to Home
+ Link to Page 3

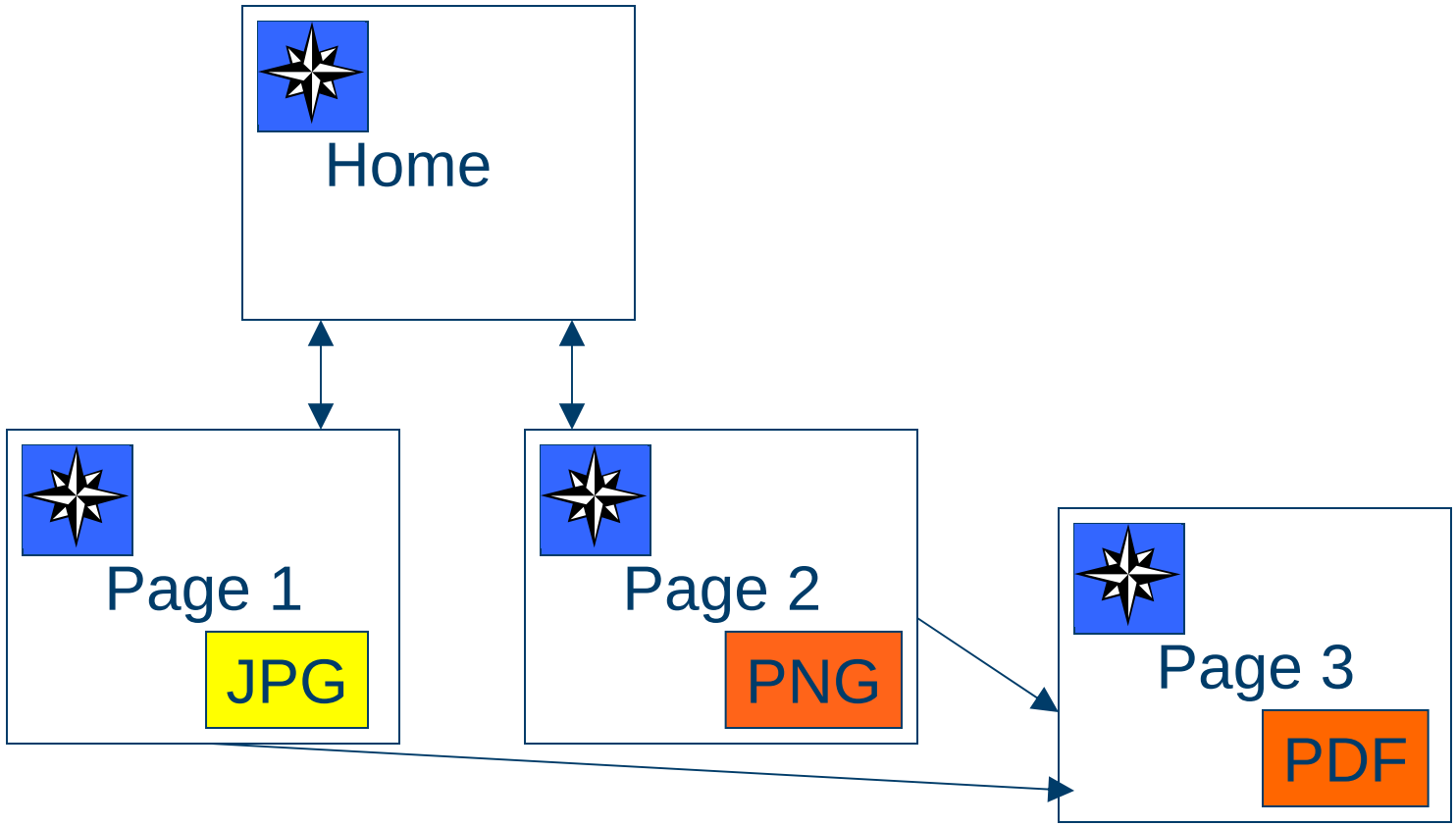
Conceptual



Tessella
Technology & Consulting

Conceptual Characterisation - Example

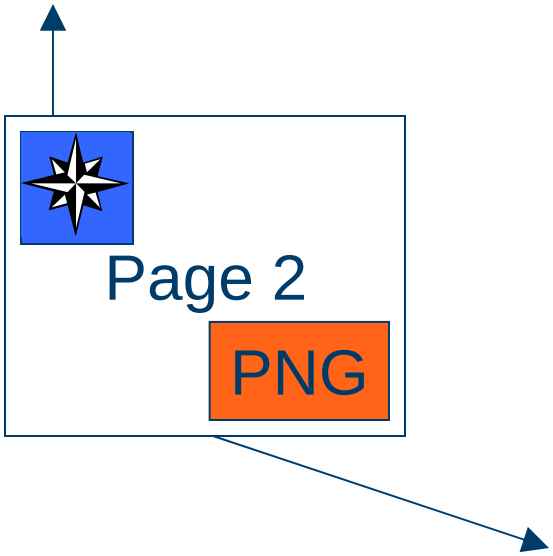
- Work out conceptual structure:



Tessella
Technology & Consulting

Conceptual Characterisation - Example

- Identify characteristics:

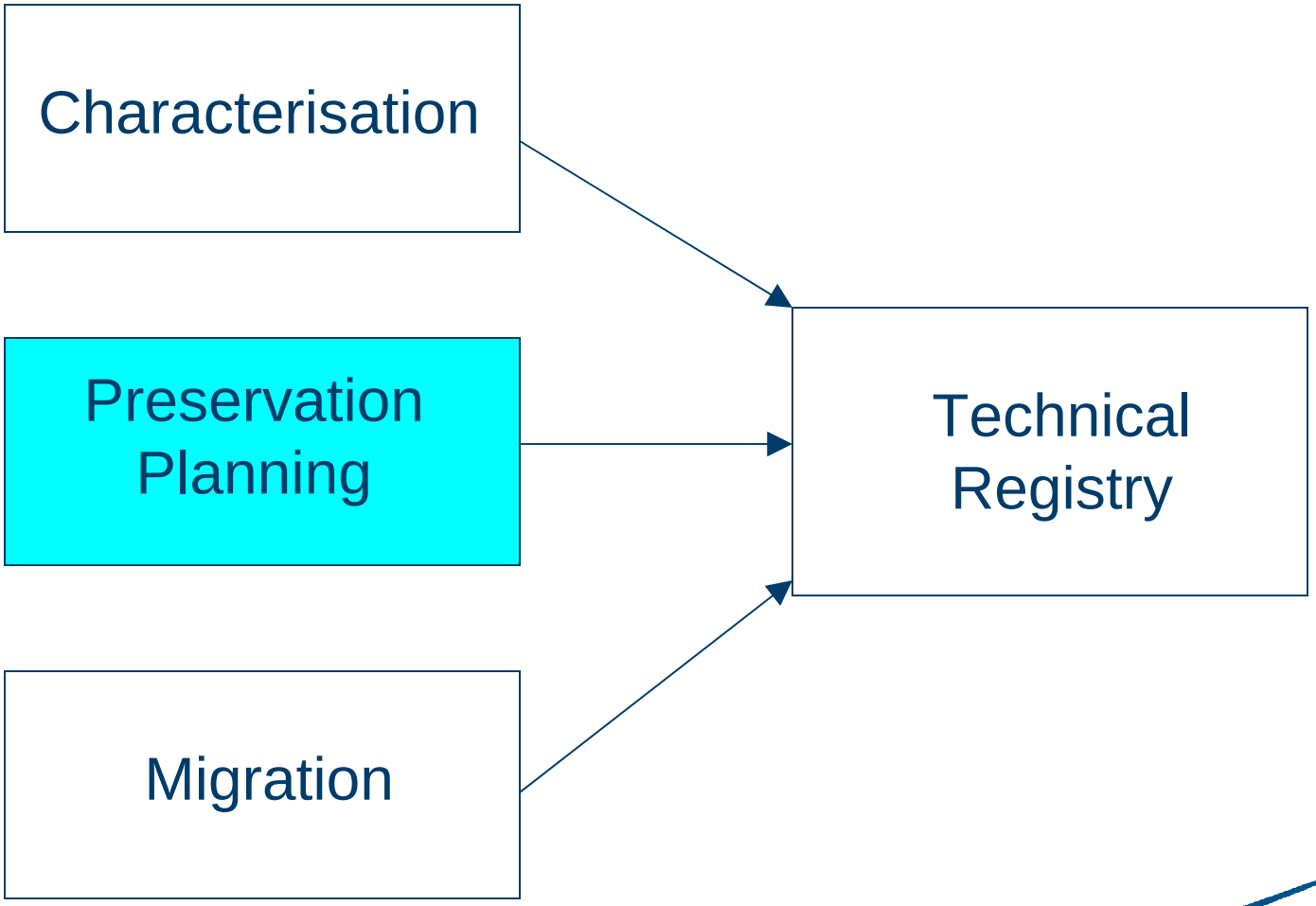


- Title
- Link to home page
- Link to Page 3
- Contains image:
  - Height
  - Width
- Contains image:
  - Height
  - Width

- Characteristics **not** linked to technology

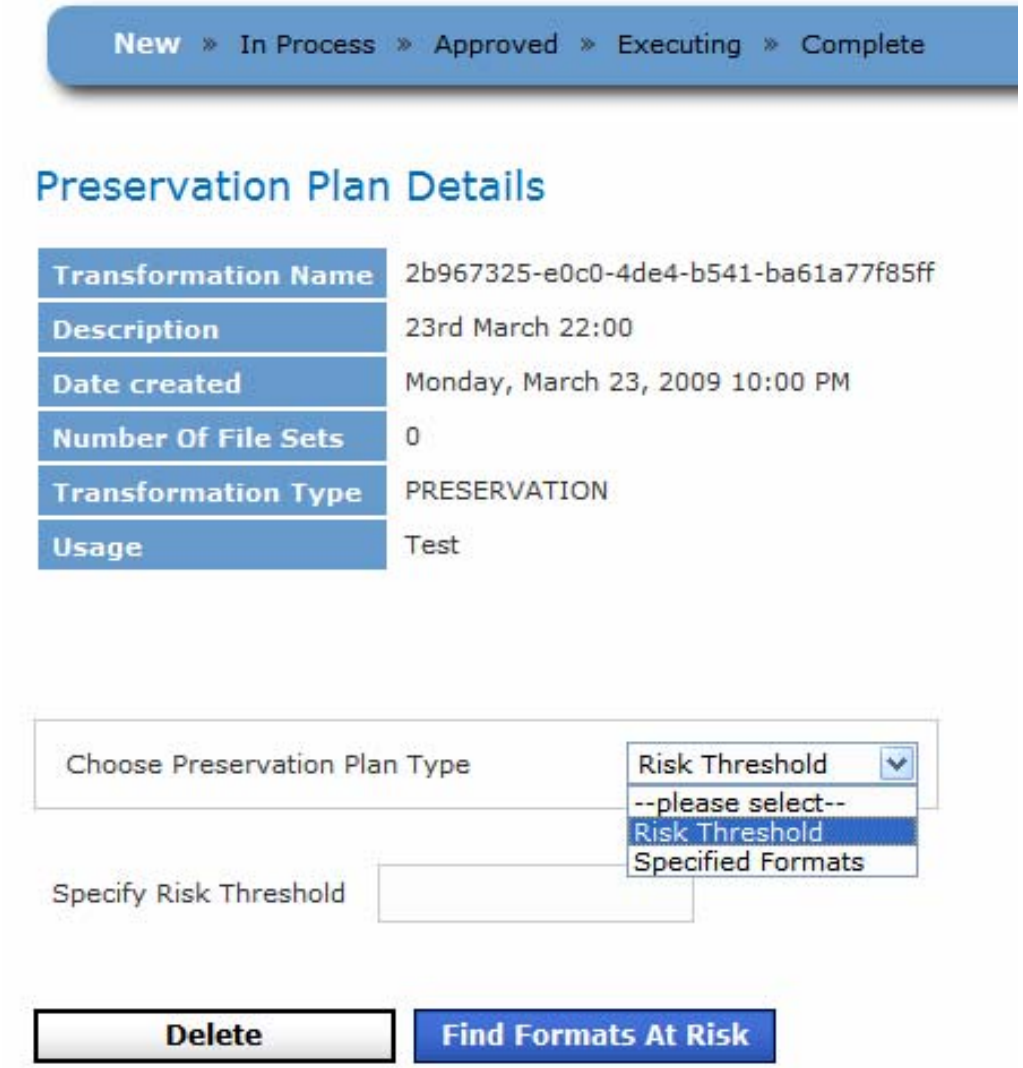
Tessella
Technology & Consulting

Active Preservation



Tessella
Technology & Consulting

Preservation Planning– Select formats



Tessella
Technology & Consulting

Preservation Planning – Risk Threshold 1/2

The following formats are considered risky at this risk threshold.

PUID Format	Format Name	Version	Risk Threshold	Options
fmt/1	Broadcast WAVE	0	Get Risk	Remove
fmt/2	Broadcast WAVE	1	Get Risk	Remove
fmt/3	Graphics Interchange Format	1987a	Get Risk	Remove
fmt/4	Graphics Interchange Format	1989a	Get Risk	Remove
fmt/5	Audio/Video Interleaved Format		Get Risk	Remove
fmt/6	Waveform Audio		Get Risk	Remove
fmt/7	Tagged Image File Format	3	Get Risk	Remove
fmt/8	Tagged Image File Format	4	Get Risk	Remove

Tessella
Technology & Consulting

Preservation Planning – Risk Threshold 2/2

The following formats do not have any available migration pathways and have been removed from the transformation.

PUID Format	Format Name	Version
fmt/1	Broadcast WAVE	0
fmt/2	Broadcast WAVE	1
fmt/5	Audio/Video Interleaved Format	
fmt/6	Waveform Audio	
fmt/11	Portable Network Graphics	1.0
fmt/12	Portable Network Graphics	1.1
fmt/13	Portable Network Graphics	1.2
fmt/21	AutoCAD Drawing	1.0
fmt/22	AutoCAD Drawing	1.2
fmt/23	AutoCAD Drawing	1.3
fmt/24	AutoCAD Drawing	1.4
fmt/25	AutoCAD Drawing	2.0
fmt/26	AutoCAD Drawing	2.1
fmt/27	AutoCAD Drawing	2.2

Tessella
Technology & Consulting

Preservation Planning – Determine Information Objects at risk…

Select the File Sets for Transformation

Select	Ingested File Set Ref	Catalogue Ref	Number of Files	Files at Risk	Size
☑	a5ec6914-f89c-4152-9ce6-f23277ea3285	Phones 1	97	32	1.43 MB
☑	4990ddba-2957-4729-9eb8-21bb114bb4b1	test0001	20	4	408.86 kB
☑	f0458fce-d1cf-42c3-9da1-a1c91bbc02fd	HP4	99	4	16.09 kB
☑	c6a04eb5-127b-4b3a-a965-3b8d8b7a4935	COLL CODE	96	90	687.17 MB
Totals	4		312	130	689.02 MB

Tessella
Technology & Consulting

Preservation Planning – Select pathways

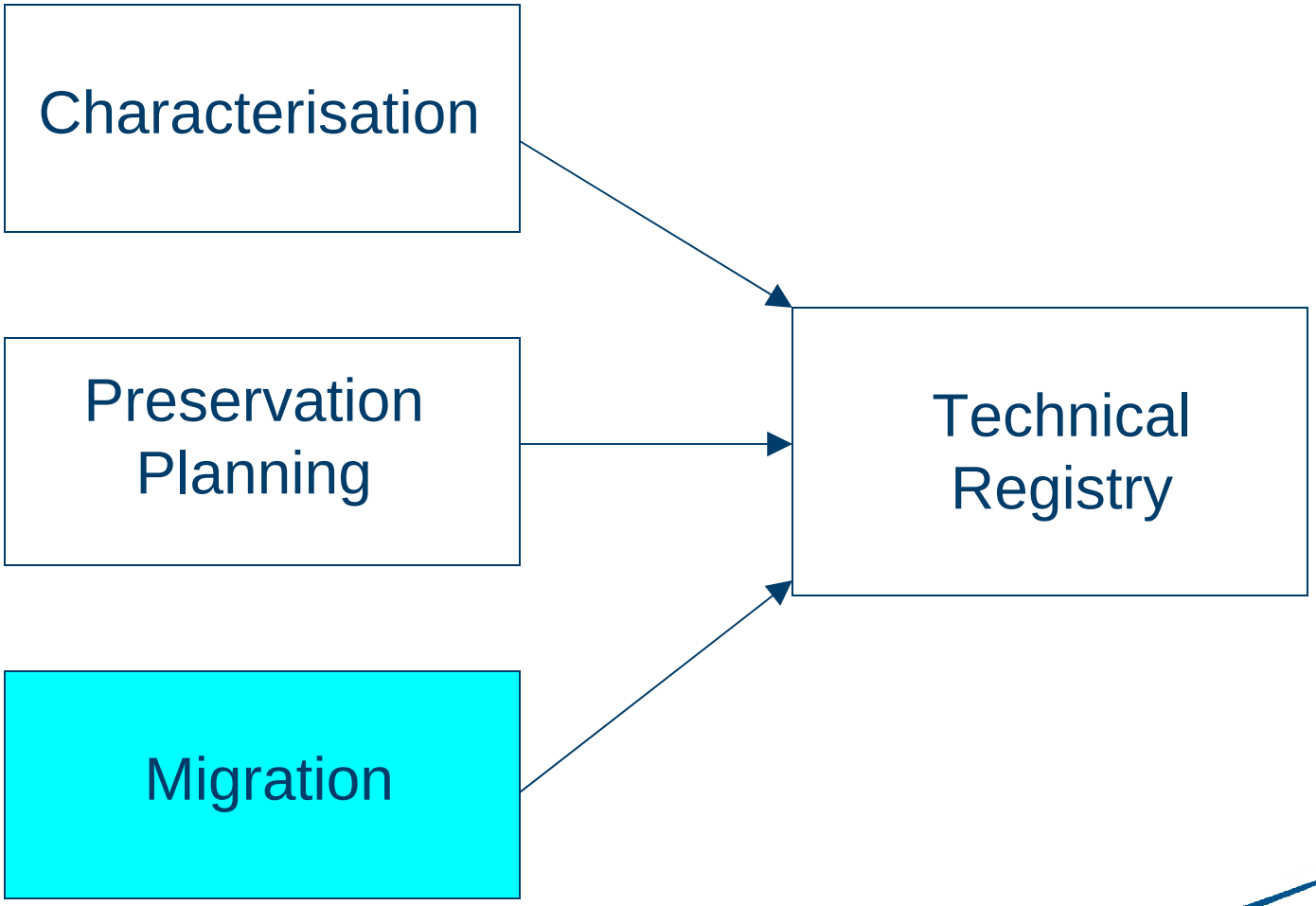
Select Transformation Pathways

Original Format	Resulting Format
JPEG File Interchange Format 1.02 (fmt/44)	Portable Network Graphics 1.0 (fmt/11) - Stellent Image Export
JPEG File Interchange Format 1.01 (fmt/43)	Tagged Image File Format 4 (fmt/8) - Image Magick
Tagged Image File Format 5 (fmt/9)	Portable Network Graphics 1.0 (fmt/11) - Stellent Image Export
Microsoft Word for Windows Document 97-2003 (fmt/40)	Portable Document Format 1.4 (fmt/18) - Stellent PDF Export
Tagged Image File Format 6 (fmt/10)	Portable Network Graphics 1.0 (fmt/11) - Stellent Image Export
Portable Document Format 1.2 (fmt/16)	Extensible Markup Language 1.0 (fmt/101) - xmlExportWrapper
Graphics Interchange Format 1989a (fmt/4)	Portable Network Graphics 1.0 (fmt/11) - Stellent Image Export

Back | Filter By Catalogue Ref | Approve

Tessella
Technology & Consulting

Active Preservation



Tessella
Technology & Consulting

Migration – Waiting…

New » In Process » Approved » **Executing** » Complete

Preservation Plan Details

Transformation Name	20cb5260-6b26-42ca-b817-b4ece740c21a
Description	23rd March 21:55
Date created	Monday, March 23, 2009 9:55 PM
Number Of File Sets	1
Transformation Type	PRESERVATION
Usage	Production

Transformation in Progress. Start Date: 23.03.09 21:57:37

Ingested File Set Ref	Number of Files	Size	Status
4990ddba-2957-4729-9eb8-21bb114bb4b1	20	399 KB	Copying Files from storage

Stop Transformation

Tessella
Technology & Consulting

Migration – Done

New » In Process » Approved » Executing » **Complete**

Preservation Plan Details

Transformation Name	20cb5260-6b26-42ca-b817-b4ece740c21a
Description	23rd March 21:55
Date created	Monday, March 23, 2009 9:55 PM
Number Of File Sets	1
Transformation Type	PRESERVATION
Usage	Production

The transformation has completed successfully.

Accession Ref	Ingested File Set Ref	New Ingested File Set Ref
7f5c5e7c-5807-4594-b609-10f919919993	4990ddba-2957-4729-9eb8-21bb114bb4b1	013e97ec-fece-4996-bfcc-be5c323fba42

Home Page

Tessella
Technology & Consulting

Migration

- Perform plan
  - Conceptual component is atomic level of migration:
    - Consume a set of files
    - Create a set of files
    - Not necessarily 1-to-1.
  - Aggregate new component manifestations to get new information object manifestation
- Verification:
  - Characterise new files
  - Re-characterise new manifestation of 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

Tessella
Technology & Consulting

Active Preservation: Summary

- Framework
  - Have tools for common formats
- Integral part of SDB
- Future:
  - Develop best practice
  - Wrap more tools: Planets, Xena, NLNZ etc.
  - Plan to be integrated to other repository systems (e.g., Fedora)

Tessella
Technology & Consulting