EPrints Preservation

What will you know after this tutorial?

- Understand the challenges in digital Preservation
- Understand why we need to plan preservation activities
- Be able to choose a simplistic preservation plan strategy.
- Be able to deploy this plan on your repository and control the outputs.

Overview

(1) Introduction

(2) File Classification & Risk Analysis

(3) Simple Preservation Planning

(4) Preservation Action & Provenance

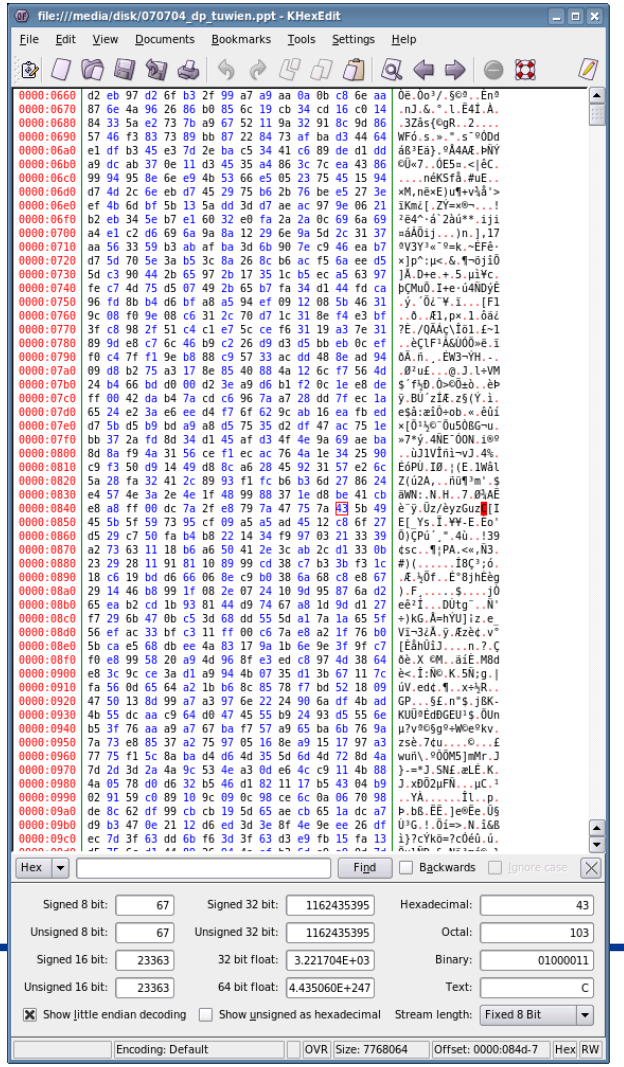
eprints

EPrints Preservation

Introduction

Why do we need Digital Preservation?

Why do we need Digital Preservation?



eprints

Why do we need Digital Preservation?

- Digital Objects require specific environment to be accessible :
  - Files need specific programs
  - Programs need specific operating systems (-versions)
  - Operating systems need specific hardware components
- SW/HW environment is not stable:
  - Files cannot be opened anymore
  - Embedded objects are no longer accessible/linked
  - Programs won‘t run
  - Information in digital form is lost
    (usually total loss, no degradation)
- Digital Preservation aims at maintaining digital objects authentically usable and accessible for long time periods.

Why do we need Digital Preservation?

- Essential for all digital objects
  - Office documents, accounting, emails, …
  - Scientific datasets, sensor data, metadata, …
  - Applications, simulations,…
- All application domains
  - Cultural heritage data
  - eGovernment, public administration
  - Science / Research
  - Industry
  - Health, pharmaceutical industry
  - Aviation, control systems, construction, …
  - Private data
  - …

Migration

- Transformation into different format, continuous or on-demand (Viewer)
  - + Wide-spread adoption
  - + Possibility to compare to un-migrated object
  - + Immediately accessible
  - - Unintended changes, specifically over sequence of migrations
  - - Cannot be used for all objects
  - - Requires continuous action to migrate

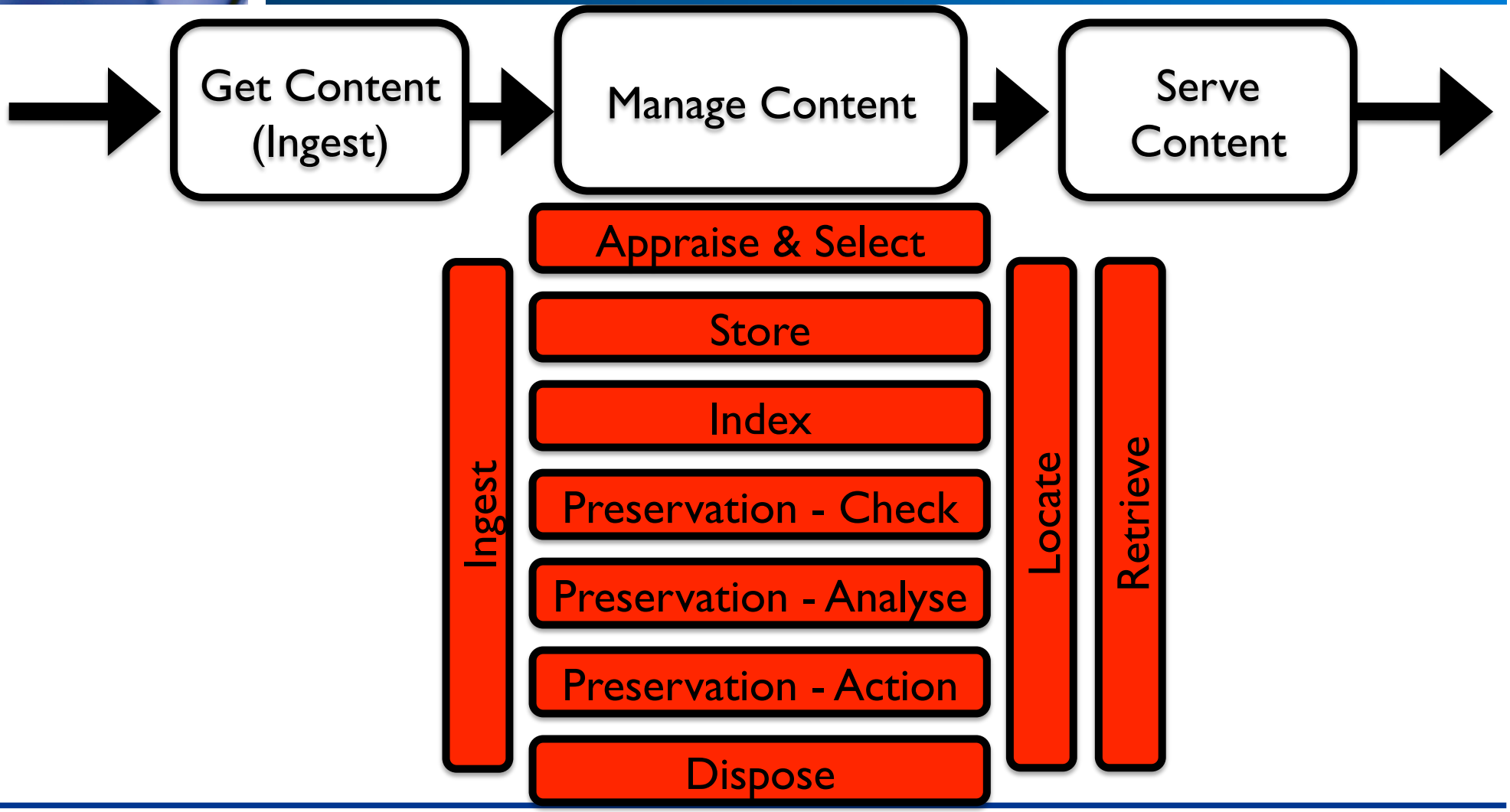
Emulation

- Emulation of hardware or software (OS, applications)
  - + Concept of emulation widely used
  - + Numerous emulators are available
  - + Potentially complete preservation of functionality
  - + *Object is rendered identically*
  - - *Object is rendered identically*
  - - Requires detailed documentation of system
  - - Requires knowledge on how to operate current systems in the future
  - - Complex technology
  - - Emulators must be emulated or migrated themselves
  - - Emulators potentially erroneous/incomplete

Open Archival Information System (OAIS) *reference* model

eprints

The 3-stage repository model



eprints

Digital Preservation

- Is a complex task
- Requires a concise understanding of the objects, their intellectual characteristics, the way they were created and used and how they will most likely be used in the future
- Requires a continuous commitment to preserve objects to avoid the “digital dark hole”
- Requires a solid, trusted infrastructure and workflows to ensure digital objects are not lost
- Is essential to maintain electronic publications & data accessible
- Will become more complex as digital objects become more complex
- Needs to be defined in a preservation plan

EPrints Preservation

File Classification & Risk Analysis

The Preservation Process

Preservation - Check

- Bit checking & checksum calculation

Preservation - Analyse

- What is the type of file, is the file valid?
- Is the file at risk of not having an editor/reader?
- Is there a better format available? Lossless or Lossy?

Preservation - Action

- File migration to avert risks found by analysis.
- Movement of file to new storage.

eprints

File Format Analysis

Preservation - Analyse

EPrints File Classification

Preserv 2

eprints

Home | About | Browse by Year | Browse by Subject

Logged in as Mr David C Tarrant | Manage deposits | Profile | Saved searches | Review | Admin | Logout

Search

Formats/Risks

Risks analysis functionality is currently not available. This feature is due to be made available by The National Archives (UK) in the near future. This page will automatically pick up the data when this feature becomes available.

No Risk Scores Available

Portable Document Format (Version 1.4)	3
Microsoft Powerpoint Presentation (Version 97-2002)	3
Portable Document Format (Version 1.3)	2
ZIP Format	2
OLE2 Compound Document Format	1

eprints

Analysis

Preservation - Analyse

- What is the type of file, is the file valid?
  - Droid is a good classification tool for this.

The National Archives

- Is the file at risk of not having an editor/reader?

- Is there a better format available? Lossless or Lossy?

Risk Information obtained from factual data
Objective risk information is local

eprints

Risk Analysis In EPrints

Preservation - Analyse

EPrints File Classification + Risk Analysis

Training Repository - 10

eprints

Home | About | Browse by Year | Browse by Subject | Browse by Division | Browse by Author

Logged in as Unnamed user with email davetaz@ecs.soton.ac.uk | Manage deposits | Manage Records | Profile | Saved searches | Review | Admin | Edit page phrases | Logout

Search

Formats/Risks

This EPrints install may be referecing a trial version of the risk analysis service. If you feel this is incorrect please contact the system administrator.

High Risk Objects

Graphics Interchange Format (Version 1987a) 9

Low Risk Objects

Acrobat PDF 1.4 - Portable Document Format (Version 1.4) 20

Graphics Interchange Format (Version 1989a) 3

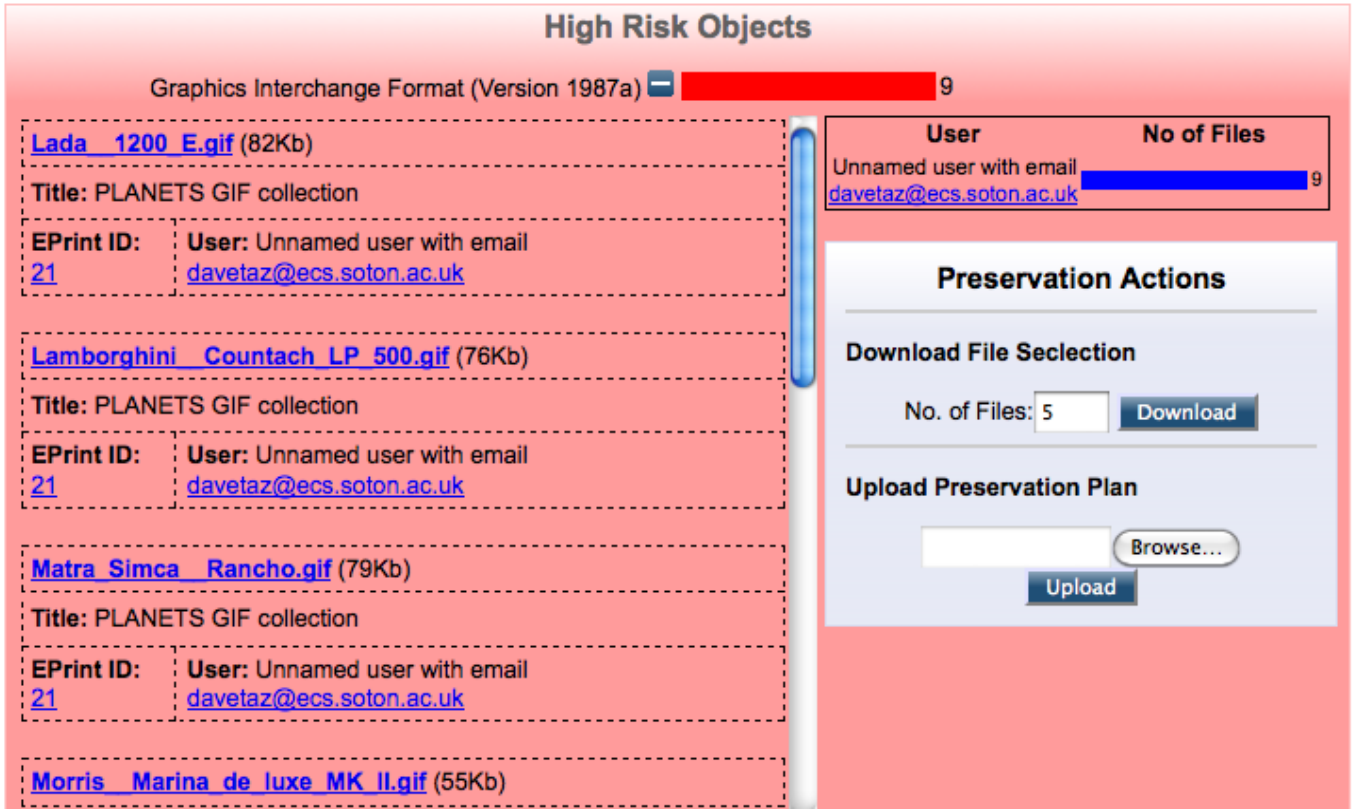
Preservation Plans

eprints

Risk Analysis In EPrints - Detailed View

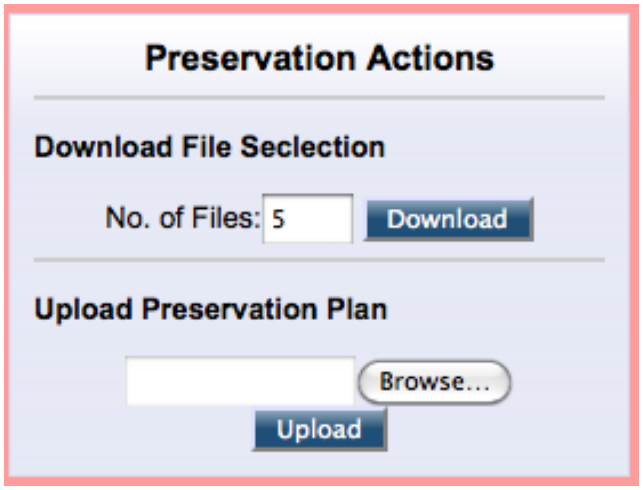
Preservation - Analyse

EPrints File Classification + Risk Analysis



eprints

Collection Gathering



- If more than 1 file requested:
  - Provide Newest and Oldest
- If morn than 3 files requested:
  - Also provide Largest and Smallest
- Then
  - Provide a random selection

Exercise Time

Recap

Preservation - Check

- Handled by our storage manager.

Preservation - Analyse

- Parallels can be drawn with storage, in that we are integrating with and utilising currently available services to perform our analysis.
- Processing of the results leads to a powerful interface which tells us many things about the repository ecosystem and it's future.

Preservation - Action

- Next part of workshop…

EPrints Preservation

Preservation Planning

Preservation workflow

Check

- Format identification, versioning
- File validation
- Virus check
- Bit checking and checksum calculation

Tools
e.g. DROID
JHOVE
FITS

Analyse

Preservation planning
Characterisation:
Significant properties and technical characteristics, **provenance**, format, **risk factors**

Risk analysis

Tools
Plato (Planets)
PRONOM (TNA)
P2 risk registry (KeepIt)
INFORM (U Illinois)
KB

Action

- Migration
- Emulation
- Storage selection

Accepted repository formats: recent survey

- What file formats do you accept? Do you convert any to a different format?
  - ALL: Accept any format.
  - Two: Convert everything to PDF, but store the source files in the background for preservation reasons.
  - Four: Mention specifically converting Word to PDF: one seeks permission from the author to do this, and uploads as Word if permission is not granted.
  - One: Mentions converting ZIP files to PDF.

Sue Ashby, University of Portsmouth Library, Summary of responses to IR questionnaire, JISC-REPOSITORIES, 18 February 2010

Format risks

1000 Ubiquity: degree of adoption of the format
1001 Support: number of tools available which can access the format
1002 Disclosure: extent to which the format documentation is publicly disclosed
1003 Document Quality: completeness of the available documentation
1004 Stability: speed and backwards-compatibility of version change
1005 Ease of Identification: ease with which the format can be identified
1006 Ease of validation: ease with which the format can be validated
1007 Lossiness: does the format use lossy compression
1008 Intellectual Property Rights: whether or not the format in encumbered by IPR
1009 Complexity: degree of content or behavioural complexity supported

From PRONOM documentation (The National Archives), July 2008

eprints

Format risks

1000 Ubiquity: degree of adoption of the format
1001 Support: number of tools available which can access the format
~~**1002** Disclosure: extent to which the format documentation is publicly disclosed~~
~~**1003** Document Quality: completeness of the available documentation~~
1004 Stability: speed and backwards-compatibility of version change
1005 Ease of Identification: ease with which the format can be identified
1006 Ease of validation: ease with which the format can be validated
1007 Lossiness: does the format use lossy compression
1008 Intellectual Property Rights: whether or not the format in encumbered by IPR
1009 Complexity: degree of content or behavioural complexity supported

From PRONOM documentation (The National Archives), July 2008

A group task on format risks

1. **Choose two formats** to compare (e.g. Word vs PDF, Word vs ODF, PDF vs XML, TIFF vs JPEG)
2. By working through the (surviving) list of format risks **select a winner** (or a draw) between your chosen formats **for each risk category** (1 point for win)
3. **Total the scores** to find an overall winning format
4. Suggest one reason why the winning format using this method may not be the one you would choose for your repository

Exercise Time

Some thoughts about formats

- Free vs open source vs open standard:

- **MS Office** – XML – open standard
- **Open Office** – free – XML - open standard
- **PDF** page representation
- **XML** generic Web format, computational

Rosenthal: Why we can relax about preservation

"Historically, the open source community has developed rendering software for almost all proprietary formats that achieve wide use

"Even the **formats which pose the greatest problems for preservation, those protected by DRM technology**, typically have open source renderers"

Format Obsolescence: Scenarios (April 29, 2007)
http://blog.dshr.org/2007/04/format-obsolescence-scenarios.html

Work with, not against, your authors and contributors

- "Preservation begins with the author"
- U. Rochester (USA) has written its own repository software IR+ to give its authors a Web-based authoring workspace
- But which applications are widely used and popular among your authors? Digital content authoring tools are typically chosen on the basis of purpose, utility, familiarity (what is provided, supported by Information Systems?) Rarely are they chosen for format or preservation.
- Authors will craft their output in the chosen application, but will often throw away that craft if asked to convert to another format
- One approach that builds on popular formats is ICE: Integrated Content Environment, which converts formats from popular content authoring tools

An image format comparison: TIFF vs JPEG 2000?

- Studies and user reports claim JPEG 2000 to be – or at least will become – the next archiving format for digital images
- The format offers new possibilities, such as streaming, and reduces storage consumption through lossless and lossy compression. Another often claimed advantage of JPEG 2000 is that the master image can possibly serve as the access copy as well, and thus replace derived compressed, low resolution access copies.

Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings, *D-Lib Magazine*, Vol15 No. 11/12, Nov/Dec 2009, http://www.dlib.org/dlib/november09/kulovits/11kulovits.html

TIFF vs JPEG 2000?

Who's for JPEG? The major players line up

1. **The National Library of the Netherlands** evaluated JPEG 2000 against uncompressed TIFF (currently used) for storage capacity, image quality, long-term sustainability, functionality. **JPEG 2000 is recommended** as future archive format.
2. **The British Library** recently moved forward to migrate their 80-terabyte newspaper collection **from TIFF to JPEG 2000**
3. **The Wellcome Library** announced they **will use JPEG 2000** for their upcoming digitization projects

Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings, *D-Lib Magazine*, Vol15 No. 11/12, Nov/Dec 2009, http://www.dlib.org/dlib/november09/kulovits/11kulovits.html

TIFF vs JPEG 2000?

- What does Plato say?

"At this point in time **not migrating** the TIFF v6 images is the best alternative."

"However, **in one year we'll look at this plan again** to see if there are more tools available and whether or not the ones we considered in this year's evaluation have been improved."

Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings, *D-Lib Magazine*, Vol15 No. 11/12, Nov/Dec 2009, http://www.dlib.org/dlib/november09/kulovits/11kulovits.html

EPrints Preservation

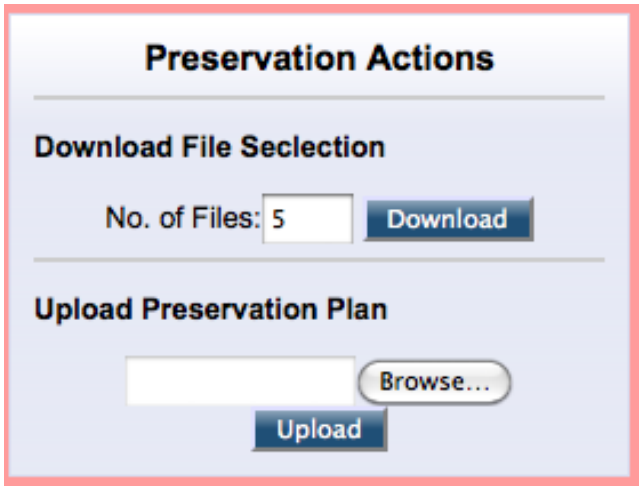
Preservation Action & Provenance

The Preservation Process

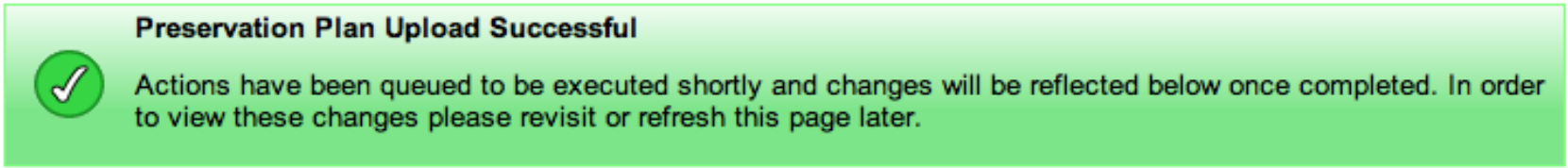
Preservation - Action

- Uploading a Preservation Plan in EPrints
- Viewing resultant actions
- Managing your plans
- Re-enacting the Plan
- Viewing Provenance Information

Uploading a Plan



- Each set of "at risk" classified files can have a single related preservation plan.
- Once uploaded, any defined actions will be performed on **all** files of that classification.

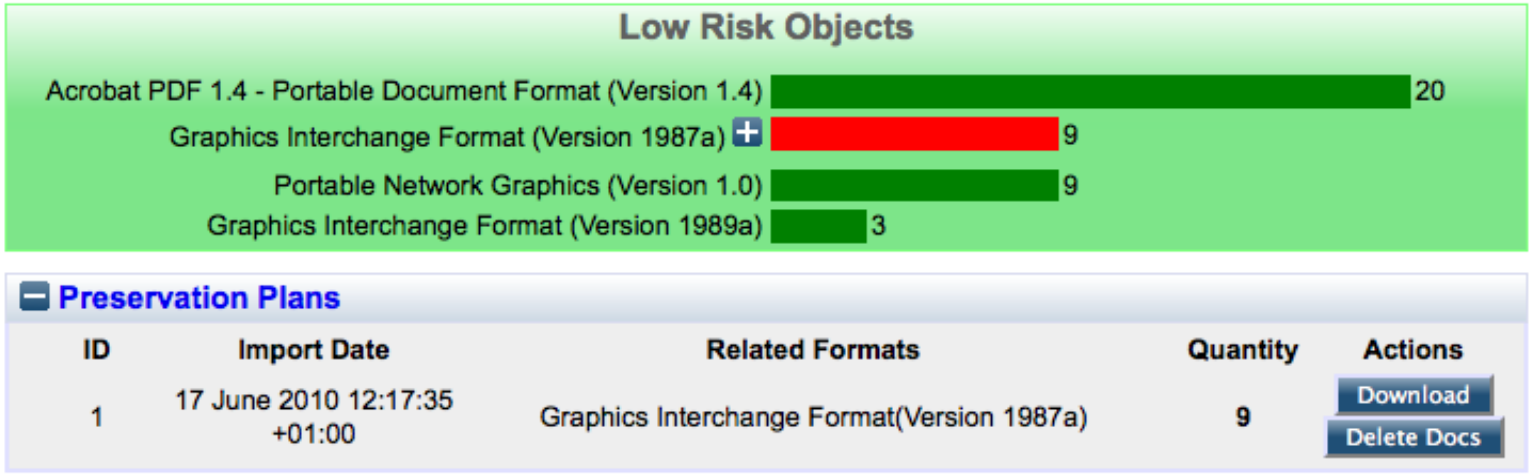


eprints

Plan Management

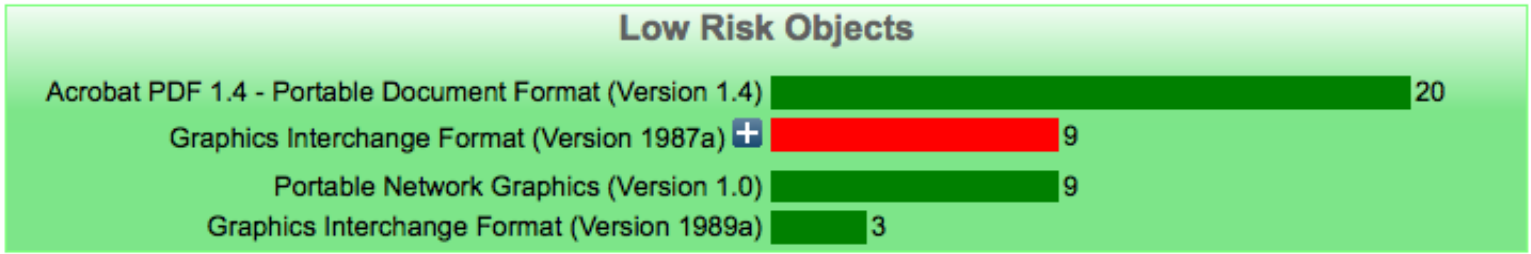
- No plan can cause files to be deleted.

- A plan controls any files it has created.
- While these files exist, the plan cannot be deleted.



Viewing the Result

- Previously high risk objects are still represented by a red bar, but are now in the low risk category.

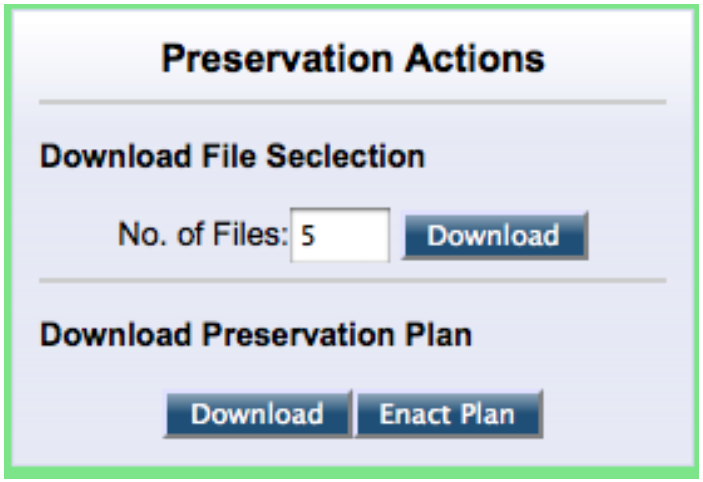


Preservation Plans

ID	Import Date	Related Formats	Quantity	Actions
1	17 June 2010 12:17:35 +01:00	Graphics Interchange Format(Version 1987a)	9	Download / Delete Docs

eprints

Preservation Actions Panel



- Download plan for reviewing in planning software.
- Re-enact plan

eprints

Viewing the Result

Before

Image (GIF)
Download (76Kb) | Preview

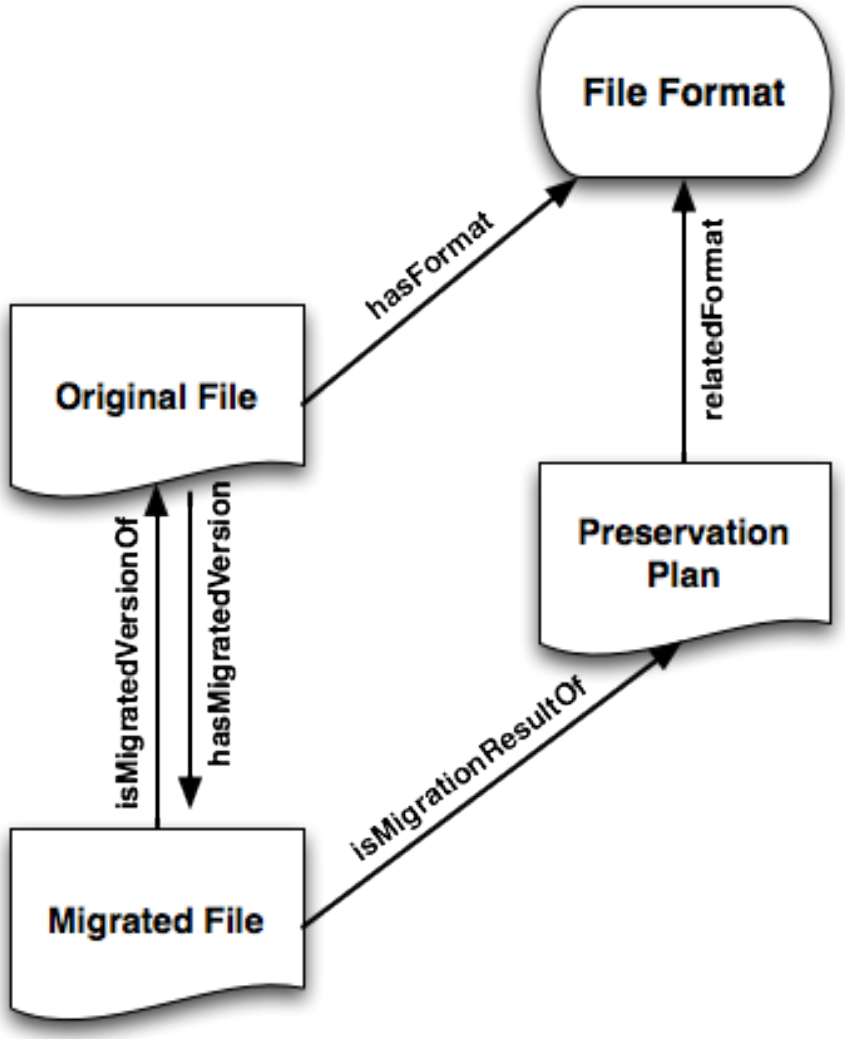
After

Image (PNG) (Migrated (Preservation) from Document ID: 41 (image/gif))
Download (76Kb) | Preview

- Image (GIF) (Original Version)

Provenance Information

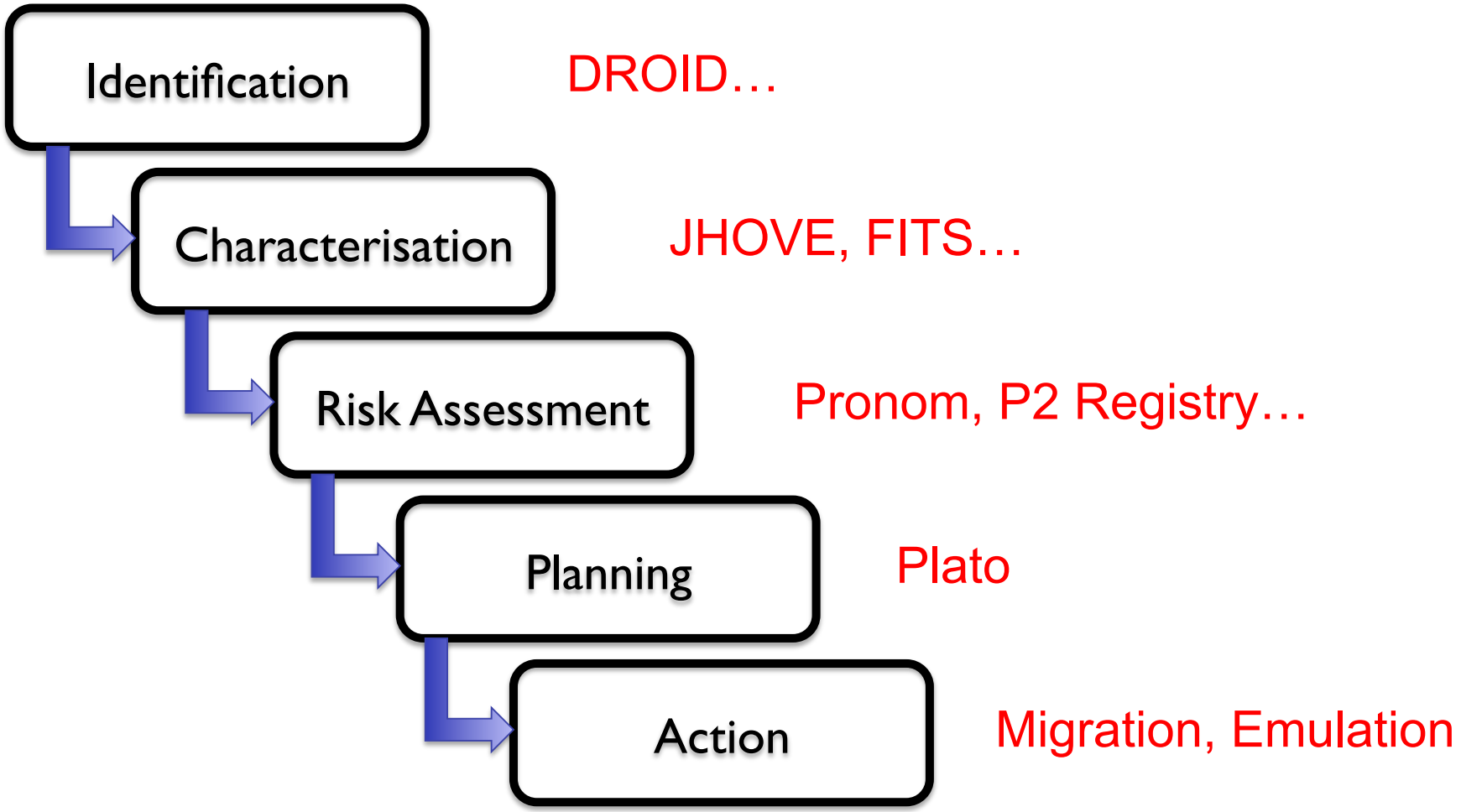
- Open Provenance Model (OPM) compliant
- Stored in RDF triple form using the EPrints relation manager added in 3.2



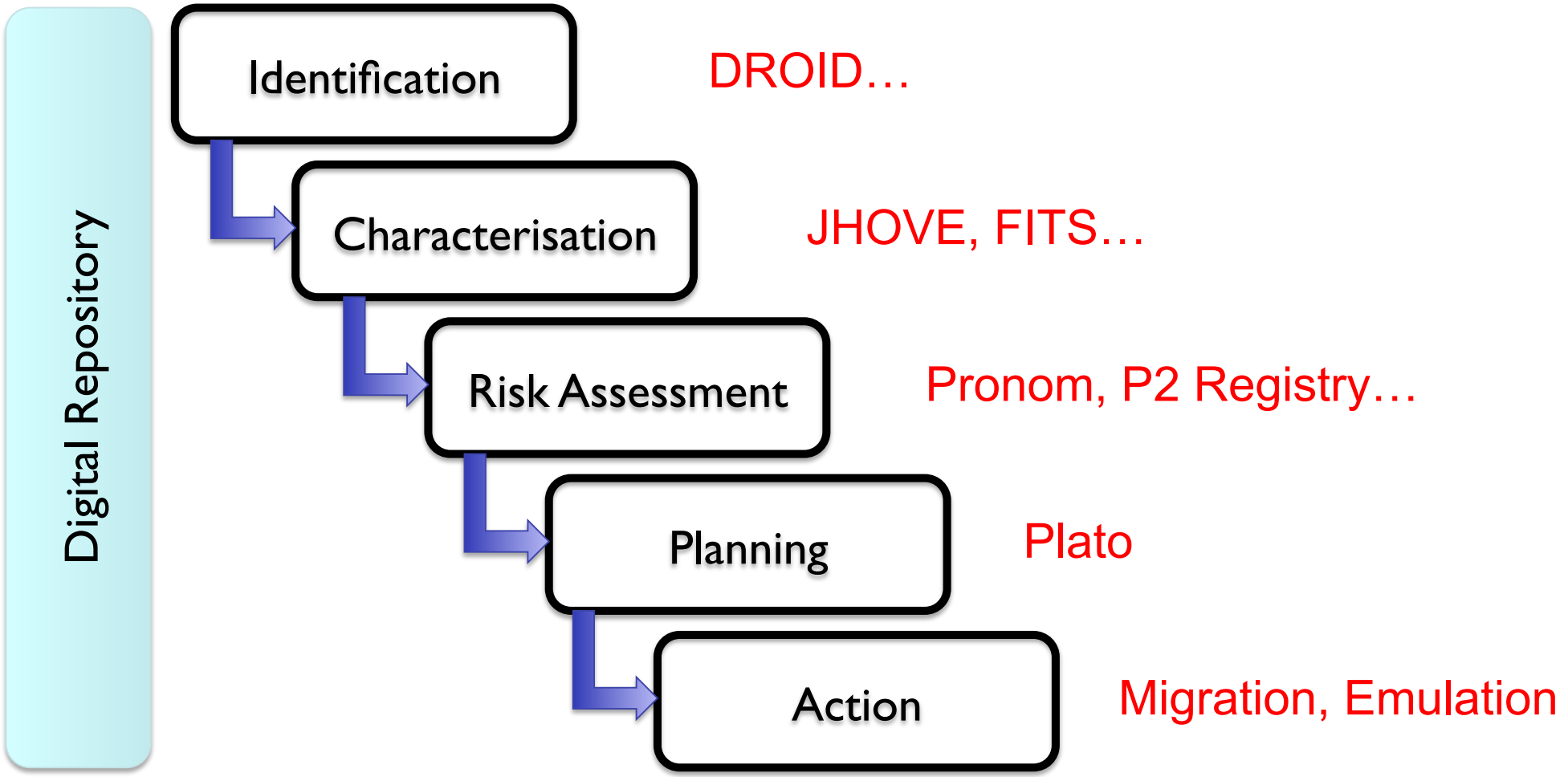
eprints

Exercise Time

Recap


Identification
DROID…
Characterisation
JHOVE, FITS…
Risk Assessment
Pronom, P2 Registry…
Planning
Plato
Action
Migration, Emulation

Recap


Digital Repository
Identification
Characterisation
Risk Assessment
Planning
Action
DROID…
JHOVE, FITS…
Pronom, P2 Registry…
Plato
Migration, Emulation

Many Thanks

planets

TU WIEN

eprints