Logical and bit-stream preservation using Plato and EPrints

Hannes Kulovits
Andreas Rauber
Department of Software Technology and Interactive Systems
Vienna University of Technology
kulovits@ifs.tuwien.ac.at
rauber@ifs.tuwien.ac.at

David Tarrant
Adam Field
School of Electronics and Computer Science
University of Southampton, UK
dct05r@ecs.soton.ac.uk
af05v@ecs.soton.ac.uk
Vienna University of Technology

- Vienna University of Technology
  http://www.tuwien.ac.at
- Faculty of Computer Science
  http://www.cs.tuwien.ac.at
  - Department of Software Technology and Interactive Systems (ISIS)
    http://www.isis.tuwien.ac.at

- People in DP
  - Andreas Rauber
  - Christoph Becker
  - Mark Guttenbrunner
  - Rudolf Mayer
  - Michael Kraxner
  - Hannes Kulovits
  - Stephan Strodl
  - Michael Greifeneder
  - Petar Petrov
DP Activities in Vienna

- Web Archiving (AOLA) in cooperation with the Austrian National Library
- DELOS DPC (EU FP6 NoE)
- DPE: Digital Preservation Europe (EU FP6 CA)
- PLANETS (EU FP6 IP)
- eGovernment & Digital Preservation series of projects with Federal Chancellery
- National Working Group on Digital Preservation of the Austrian Computer Society, in cooperation with ONB
- Digital Memory Engineering: National research studio
University of Southampton, UK

- University of Southampton
  - http://www.soton.ac.uk
    - School of Electronics & Computer Science
      - http://www.ecs.soton.ac.uk
- EPrints
  - http://www.epints.org
- People in Preservation
  - Steve Hitchcock
  - David Tarrant
  - Chris Gutteridge
  - Tim Brody
  - Patrick McSweeney
- EPrints Services
  - Adam Field
  - Tim Miles-Board
DP Activities in Southampton

- EPrints Preservation
  - KeepIt!
  - Preserv2
  - Preserv

- P2N – Preservation Network
  - Collaboration with Oxford University

- P2-Registry
  - Linked Data for Digital Preservation

- Web Archiving
  - ECS project to archive old project websites and Wikis
Introductions
What will you know after this tutorial?

You will:

- See the (first?) system integrating **bit stream preservation** and **logical preservation** supported by a **fully documented planning** process
- Perform **risk analysis as trigger** for preservation actions
- Understand why we need to **plan preservation activities**
- Know a **workflow to evaluate preservation strategies**
- Be familiar with Plato and EPrints
- Be able to **develop a specific preservation plan** that is optimized for
  - the objects in your institution
  - the users of your institution
  - the institutional requirements
- Be able to **execute it** in a repository (EPrints)
Integrated Preservation Cycle

- EPrints Repository
  - Re-Evaluation
  - Risk Analysis
- Plato
  - Preservation Planning
  - Plan Enactment
Schedule

09:00 – 09:45 Introduction
09:45 – 11:00 Exercise 1 (EPrints)
11:00 – 11:15 Coffee/Tea
11:15 – 13:00 Requirements
13:00 – 14:00 Lunch
14:00 – 15:30 Evaluation/Transformation
15:30 – 16:00 Coffee/Tea
16:00 – 17:15 EPrints
17:15 – 18:00 Discussion
(18:15 - ??? Ice breaking & Wine tasting)
Schedule

(1) Introduction

(2) Preservation in EPrints

(3) Preservation Planning with Plato

(4) Bringing it all together and Closing
Overview

Part 1: Introduction

- Quick introduction to physical preservation with EPrints
- Quick introduction to logical preservation with Plato
- Bringing it together: bit-stream and logical preservation
What is EPrints For?

- EPrints offers a safe, open and useful place to store, share and manage material in the pursuit of research and educational agendas.

administrative reporting, collaboration, data sharing, digital profile enhancement, e-learning, e-publishing, e-research, marketing, open access, preservation, publicity, research assessment, research management, scholarly collections
An EPrints repository is

- A valuable part of the researcher’s information environment
  - directly integrating with the research desktop
  - offering sustainable storage and open access

- A competent and mature component of the institution’s information environment
  - providing management and curation support for core business research data
  - leveraging information about research outputs to inform management strategy
KeepIT Exemplars

Open Access to Research Outputs

Open Educational Resources

Open Arts

Open Scientific Data
EPrints Repositories

- eprints.lse.ac.uk (institutional)
- eprints.ecs.soton.ac.uk (departmental)
- pubs.or08.ecs.soton.ac.uk (conference)
- archive.serpentproject.com (project)
- nora.nerc.ac.uk (funders)
- ecrystals.chem.soton.ac.uk (data)
- www.linnean-online.org (collection)
- ualresearchonline.arts.ac.uk (art)
- demoprints.eprints.org (demo)
Overview

Part 1: Introduction

- Quick introduction to physical preservation with EPrints
- Quick introduction to logical preservation with Plato
- Bringing it together: bit-stream and logical preservation
Preservation Planning

Why Preservation Planning?

- Several preservation strategies developed
  - For each strategy: several tools available
    - For each tool: several parameter settings available
- How do you know which one is most suitable?
- What are the needs of your users? Now? In the future?
- Which aspects of an object do you want to preserve?
- What are the requirements?
- How to prove in 10, 20, 50, 100 years, that the decision was correct / acceptable at the time it was made?
Preservation Planning

- Consistent workflow leading to a preservation plan
- Analyses, which solution to adopt
- Considers
  - preservation policies
  - legal obligations
  - organisational and technical constraints
  - user requirements and preservation goals
- Describes the
  - preservation context
  - evaluated preservation strategies
  - resulting decision including the reasoning

- Repeatable, solid evidence
Digital Preservation

What is a preservation plan?

- 10 Sections
  - Identification
  - Status
  - Description of Institutional Setting
  - Description of Collection
  - Requirements for Preservation
  - Evidence for Preservation Strategy
  - Cost
  - Trigger for Re-evaluation
  - Roles and Responsibilities
  - Preservation Action Plan

[Preservation Plan Template]
Preservation Planning

Preservation Planning in Plato

Define requirements
- Define basis
- Choose records
- Identify requirements

Evaluate alternatives
- Go/No-Go
  - Define alternatives
    - Develop experiment
    - Run experiment
    - Evaluate experiment

Consider results
- Analyse results
  - Set importance factors
  - Transform measured values

Preservation Action Recommendations
- Preservation action recommendation

Build preservation plan
- Create executable preservation plan
- Define preservation plan
- Validate preservation plan

Preservation Plan

Administration
- Proposals
  - Recommendations
- Inventory reports
  - Performance info
  - Consumer comments

Knowledge base
- Develop Preservation Strategies and Standards
  - Technology alerts
  - External data standards
  - Prototype results
  - Reports

Monitor Designated Community
- Reports
  - Requirement alerts
  - Emerging standards

Monitor Technology
- Prototype requests
  - Prototype requests

PRODUCER

TU WIEN
planets
UNIVERSITY OF SOUTHAMPTON ejprints
Preservation Planning Workflow
Identify requirements

Analog...

... or born digital
Overview

Part 1: Introduction

- Quick introduction to physical preservation with EPrints
- Quick introduction to logical preservation with Plato
- Bringing it together: bit-stream and logical preservation
Bringing it all together

- Identification
- Risk Analysis
- Characterisation
- Preservation Planning
- Bit Stream Preservation

Repository

- Storage Controller
- Droid
- Pronom
- JOVE
- Plato
- e.g. ImageMagick

TU Wien

planets

UNIVERSITY OF Southampton

ejprints
Bringing it all together (3/2)

EPrints Repository

Risk Analysis

Re-Evaluation

Plan Enactment

Preservation Planning

Plato
Conclusions

- Integrating bit-stream and logical preservation
- Thorough planning process
- Actionable preservation plan
- Consistent with OAIS model
- Follows recommendations of TRAC and nestor
- Generic workflow that can easily be integrated in different institutional settings

**EPrints:**
- Open-source repository system
  
  [http://www.eprints.org](http://www.eprints.org)

**Plato:**
- Tool support for preservation planning
  
  [http://www.ifs.tuwien.ac.at/dp](http://www.ifs.tuwien.ac.at/dp)
  [http://www.ifs.tuwien.ac.at/dp/plato](http://www.ifs.tuwien.ac.at/dp/plato)
Schedule

(1) Introduction

(2) Preservation in EPrints

(3) Preservation Planning with Plato

(4) Bringing it all together and Closing
The Preservation Process

**Preservation - Check**
- Resilient Storage
- 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 - Planning**
- What is the best preservation action given requirements and constraints
- Preservation Planning (Plato)

**Preservation - Action**
- File migration to avert risks found by analysis.
- Movement of file to new storage.
The Storage Ecosystem

Local
- No local bandwidth costs
- Hard to expand
- Locally Managed
- High overheads cost
- Requires space and cooling
- Tied closely to the software

Archival
- Specialist
- Expensive to purchase
- Locally Managed
- Space and running costs
- Expandable

Cloud
- Scalable
- Externally controlled
- Known Costings
- Unclear retention policy
- Re-Useable (APIs)
- Global Scale
Hybrid Storage

- Use the best features of each storage type
- Performance
  - Scaling-up bandwidth
- Optimisation
  - Large-file handling
  - Multimedia streaming
- Localised Delivery
  - Local delivery from the cloud
The storage controller manages the location of files. It uses rule based policy defined by a simple configuration file (XML). Examples:

- Large binary files of scientific data (raw machine result data) can be stored in a large disk (slower access) system and sent to a tape company for long term storage.
- Processed results can be stored locally and in the cloud ready for rapid delivery to end points.
Hybrid Storage Policies
EPrints Storage Manager

Amazon S3 storage
There are 217 total files stored using this back-end, taking 3126Kb.
Documents: 217

Local disk storage
There are 289 total files stored using this back-end, taking 1649Kb.
History: 289

Compressed local disk storage
There are 85 total files stored using this back-end, taking 293Kb.
History: 85

Copy to
Delete Copies

Copy to
Delete Copies

Compressed local disk storage
Amazon S3 storage
Recap

1. Storage Ecosystem
   - There are a great number of products and services available designed to protect your resources. Each is aimed at a market with different needs based on the type of content.

2. Storage Controller
   - Allows you to utilise a diverse range of storage services simultaneously. Take advantage of the current ecosystem.

3. Managing Stored Assets
   - If the ecosystem changes, moving of resources to a new service is a seamless operation.
The Preservation Process

**Preservation - Check**
- Resilient Storage
- 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.
Analysis

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

- Is the file at risk of not having an editor/reader?
  - Functionality is being developed in PRONOM technical registry.

- Is there a better format available? Lossless or Lossy?
File Format Analysis

**Preservation - Analyse**

EPrints File Classification

**Preserv 2**

**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
- Portable Document Format (Version 1.3): 2
- ZIP Format: 2
- OLE2 Compound Document Format: 1
Risk Analysis

Preservation - Analyse

- Is the file at risk of not having an editor/reader?
  - Functionality is being developed in PRONOM technical registry.
- Simple SOAP web service
- Takes file format identification id’s, hands back risk score.
- Breakdown of risk score may also be available in future releases.
- A stub you can download and run providing this functionality before the official release with mock up risk scores is available at http://preserv2.googlecode.com
Risk Analysis In EPrints

Preservation - Analyse

EPrints File Classification + Risk Analysis

Preserv 2

Formats/Risks

This EPrints install is referencing a trial version of the risk analysis service. None of the risk scores are likely to be accurate and thus should not be used as the basis for a program of action.

High Risk Objects

- OLE2 Compound Document Format

Medium Risk Objects

- Microsoft Powerpoint Presentation (Version 97-2002)

Low Risk Objects

- Portable Document Format (Version 1.4)
- Portable Document Format (Version 1.3)
- ZIP Format
### High Risk Objects

<table>
<thead>
<tr>
<th>File Name</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lada_1200_E.gif</td>
<td>82Kb</td>
<td>PLANETS GIF collection</td>
</tr>
<tr>
<td>Title: PLANETS GIF collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPrint ID: 21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>User: Unnamed user with email</td>
<td></td>
<td><a href="mailto:cavetaz@ecs.soton.ac.uk">cavetaz@ecs.soton.ac.uk</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamborghini_Countach_LP_500.gif</td>
<td>76Kb</td>
<td>PLANETS GIF collection</td>
</tr>
<tr>
<td>Title: PLANETS GIF collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPrint ID: 21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>User: Unnamed user with email</td>
<td></td>
<td><a href="mailto:cavetaz@ecs.soton.ac.uk">cavetaz@ecs.soton.ac.uk</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matra_Simca_Rancho.gif</td>
<td>79Kb</td>
<td>PLANETS GIF collection</td>
</tr>
<tr>
<td>Title: PLANETS GIF collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPrint ID: 21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>User: Unnamed user with email</td>
<td></td>
<td><a href="mailto:cavetaz@ecs.soton.ac.uk">cavetaz@ecs.soton.ac.uk</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morris_Marina_de_luxe_MK_Il.gif</td>
<td>55Kb</td>
<td>PLANETS GIF collection</td>
</tr>
<tr>
<td>Title: PLANETS GIF collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPrint ID: 21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>User: Unnamed user with email</td>
<td></td>
<td><a href="mailto:cavetaz@ecs.soton.ac.uk">cavetaz@ecs.soton.ac.uk</a></td>
</tr>
</tbody>
</table>

**Preservation Actions**

- **Download File Selection**
  - No. of Files: 5
  - Download

- **Upload Preservation Plan**
  - Browse... 
  - Upload
Exercise Time
Recap

Preservation - Check

- Handled by our storage manager and reported back via the preservation interface.

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

- Future plan is to utilise further web based services to ensure information remains comprehensive and up to date set, 0day digital preservation.
Schedule

(1) Introduction
   - EPrints
   - Preservation Planning and Plato

(2) Preservation in EPrints

(3) Preservation Planning with Plato

(4) Bringing it all together and Closing
Overview

Part 3: Preservation Planning with Plato

- Preservation planning workflow
- Exercises
PP Workflow

Preservation Planning in Plato

Define requirements

Define basis → Choose records → Identify requirements

Evaluate alternatives

Go/No-Go → Define alternatives

Develop experiment → Run experiment → Evaluate experiment

Consider results

Analyse results → Set importance factors → Transform measured values

Preservation Action Recommendation

Build preservation plan

Create executable preservation plan → Define preservation plan → Validate preservation plan

Preservation Plan

Knowledge base

Tree templates and fragments

Mapping characteristics to requirements
Orientation

Preservation Planning in Plato

Define requirements
- Define basis
- Choose records
- Identify requirements

Evaluate alternatives
- Go/No-Go
- Define alternatives
- Develop experiment
- Run experiment
- Evaluate experiment

Consider results
- Analyse results
- Set importance factors
- Transform measured values

Preservation Action Recommendation

Build preservation plan
- Create executable preservation plan
- Define preservation plan
- Validate preservation plan

Preservation Plan

Tree templates and fragments
Mapping characteristics to requirements
Knowledge base

TU WIEJN
planets
UNIVERSITY OF SOUTHAMPTON
 ejprints
Define Basis

- Basic preservation plan properties
- Describe the context
  - Institutional settings
  - Legal obligations
  - User groups, target community
  - Organisational constraints
- 5 triggers
  - New Collection Alert (NCA)
  - Changed Collection Profile Alert (CPA)
  - Changed Environment Alert (CEA)
  - Changed Objective Alert (COA)
  - Periodic Review Alert (PRA)
Define Basis

Organizational structure

- Mandate, Mission Statement
  - Provide reliable, long-term access to digital objects
  - Internet Archive: “The Internet Archive is working to prevent the Internet [...] and other ‘born digital’ materials from disappearing into the past. Collaborating with institutions including the Library of Congress and the Smithsonian, we are working to preserve a record for generations to come.”
    http://www.archive.org/about/about.php
  - Oxford Digital Library: “Like traditional collection development long-term sustainability and permanent availability are major goals for the Oxford Digital Library.”
    http://www.odl.ox.ac.uk/principles.htm
Orientation
Choose Sample Objects

- Identify consistent (sub-)collections
  - Homogeneous type of objects (format, use)
  - To be handled with a specific (set of) tools

- Describe the collection
  - What types of objects?
  - How many?
  - Which format(s)?

- Selection
  - Representative for the objects in the collection
  - Right choice of sample is essential
  - They should cover all essential features and characteristics of the collection in question
  - As few as possible, as many as needed
  - Often between 3 – 10
Choose Sample Objects

- Stratification – all essential groups of digital objects should be chosen according to their relevance

- Possible stratification strategies
  - File type
  - Size
  - Content (e.g. document with lots of images, including macros)
  - Time (objects from different periods of times)

- File Format Identification
  - DROID
  - PRONOM
Define Sample Objects

**Sample Records**

<table>
<thead>
<tr>
<th>Sample Record</th>
<th>Object Format</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full name:</strong></td>
<td>PUID:</td>
</tr>
<tr>
<td>sample thesis 1</td>
<td></td>
</tr>
<tr>
<td><strong>Short name:</strong></td>
<td>Name:</td>
</tr>
<tr>
<td>DA1</td>
<td></td>
</tr>
<tr>
<td><strong>Has data:</strong></td>
<td>Version:</td>
</tr>
<tr>
<td>✓ download</td>
<td></td>
</tr>
<tr>
<td>Original technical environment:</td>
<td>Mime-type:</td>
</tr>
<tr>
<td>Description:</td>
<td>Identify format</td>
</tr>
</tbody>
</table>

Remove record

Add new record without file

Add record

Upload new record

Save  Discard changes  Proceed
Practise time!

- Public institution – State and University Library
- Mission to preserve the state’s cultural heritage in the form of any publication
- Scanned collection of yearbooks, 9000 objects
  - One file per page
  - Scans are black and white
- Copyright held for the physical material, same for digital content
- Objects are provided
Orientation
Identify Requirements

- Define all relevant goals and characteristics (high-level, detail) with respect to a given application domain
- Put the requirements in relation to each other → Tree structure
- Top-down or bottom-up
  - Start from high-level goals and break down to specific criteria
  - Collect criteria and organize in tree structure
Input needed from a wide range of persons, depending on the institutional context and the collection.
Identify requirements

- Core step in the process
- Define all relevant goals and characteristics (high-level, detail) with respect to given application domain
- Usually four major groups
  - Object characteristics (content, metadata,…)
  - Record characteristics (context, relations,…)
  - Process characteristics (scalability, error-detection,…)
  - Costs (set-up, per object, HW/SW; personnel,…)

![Diagram of Preservation Planning in Pons](attachment:image.png)
Identify requirements

analogue...

... or digital
Identify requirements

Example: Webarchive
Identify requirements

- Creation within PLATO with Tree-Editor
Identify requirements

- Assign measurable unit to each leaf criterion
  - As far as possible automatically measurable
    - seconds / Euro per object
    - colour depth in bits
    - ...
  - Subjective measurement units where necessary
    - diffusion of file format
    - amount of expected support
    - ...
  - No limitations on the type of scale used
Identify requirements

Types of scales
- Numeric
- Yes/No (Y/N)
- Yes/Acceptable/No (Y/A/N)
- Ordinal: define the possible values
- Subjective 0-to-5
Identify requirements

- Creation within PLATO with Tree-Editor

---

**PLANETS Preservation Planning Tool (Plato)**

**Identify Requirements**

**Objective Tree**

**Descriptive Information**

**How can I define the objective tree?**

<table>
<thead>
<tr>
<th>Focus</th>
<th>Node</th>
<th>Single</th>
<th>Scale</th>
<th>Restriction</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Website</td>
<td>Record characteristics</td>
<td></td>
<td>Ordinal</td>
<td>Usurpative/Widespread/Special</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Technical characteristics</td>
<td></td>
<td>Positive Integer</td>
<td></td>
<td>number of tools</td>
</tr>
<tr>
<td></td>
<td>Ubiquity</td>
<td></td>
<td>Ordinal</td>
<td>Automatic/Manual/No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Support</td>
<td></td>
<td>Ordinal</td>
<td>Automatic/Manual/No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Documentation</td>
<td></td>
<td>Ordinal</td>
<td>Lossy/Lossless</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stability</td>
<td></td>
<td>Boolean</td>
<td>Yes/No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ease of identification</td>
<td></td>
<td>Ordinal</td>
<td>Automatic/Manual/No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ease of validation</td>
<td></td>
<td>Ordinal</td>
<td>High/Medium/Low</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lossiness</td>
<td></td>
<td>Ordinal</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IPR</td>
<td></td>
<td>Boolean</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complexity</td>
<td></td>
<td>Ordinal</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**TU WIEN**
Example Webarchiving:
- Static Webpages
- Including linked documents such as doc, pdf
- Images
- Interactive elements need not be preserved
Identify Requirements: Example

- Ubiquitous/Widespread/Specialised/Obsolete
- Primary/Secondary
- Full/Partial/None
- Standard/Open/Proprietary
- Public/Limited/None
- <1 year/1-2 years/3-5 years/>5 years
- None/Previous version only/Some previous versions/All previous versions
- Automatic/Manual/No
- Lossy/Lossless
- High/Medium/Low
- None/Detectable/Recoverable
- Larger/Same/Smaller

- Ubiquity
- Support
- Quality
- Disclosure
- Openness
- Availability
- Speed of change
- Backwards compatibility
- Stability
- Ease of identification
- Ease of validation
- Lossiness
- IPR
- Complexity
- Error tolerance
- Comparative size
Visitor counter and similar functionalities can be

- Frozen at harvesting time
- Omitted
- Remain operational, i.e. the counter will be increased upon archival calls
  (is this desired? count? demonstrate functionality?)
Practise time!

- Go to Plato: [http://www.ifs.tuwien.ac.at/dp/plato](http://www.ifs.tuwien.ac.at/dp/plato)
- Log into Plato with group account
- Click “List my preservation plans”

- Open preservation plan named
  
  “Scanned yearbooks archive (IDENTIFY REQUIREMENTS)”

- Enter further requirements
PP Workflow
Orientation
Define Alternatives

- Given the type of object and requirements, what strategies are possible and which is most suitable
  - Migration, emulation, other?

- For each alternative, precise definition of
  - Which tool (OS, version)
  - Which functions of the tool
  - Which parameters
  - Resources that are needed (human, technical, time and cost)

- Define manually or use registries via web services
Define Alternatives

PLANETS Preservation Planning Tool (Plato)

Define the alternatives of the Project

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>196616</td>
<td>TIFF (tool A)</td>
<td>Convert to TIFF using the well-tested and expensive tool 'A'</td>
</tr>
<tr>
<td>196613</td>
<td>TIFF (tool B)</td>
<td>Convert to TIFF/4 using this new tool named 'B'</td>
</tr>
<tr>
<td>196614</td>
<td>GIF (tool C)</td>
<td>Convert to GIF using the well-tested tool 'C'</td>
</tr>
<tr>
<td>196615</td>
<td>PNG (tool D)</td>
<td>Convert to PNG using the well-tested tool 'D'</td>
</tr>
</tbody>
</table>

Add new Alternative
Save Discard changes Proceed

Create alternatives from applicable services

Sample record #1 has format JPEG File Interchange Format, 1.01.
You can look up services that are able to handle this object type in the following registries:

<table>
<thead>
<tr>
<th>Preservation Action</th>
<th>Target Format</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPG &gt; BMP</td>
<td>Windows Bitmap, version 3.0</td>
<td>JPG&gt;BMP</td>
</tr>
<tr>
<td>JPG &gt; TIFF</td>
<td>Tagged Image File Format, version 3</td>
<td>JPG&gt;BMP&gt;TIFF</td>
</tr>
<tr>
<td>JPG &gt; TIF #2</td>
<td>Tagged Image File Format, version 3</td>
<td>JPG&gt;TIF</td>
</tr>
<tr>
<td>JPG &gt; TIF_2</td>
<td>Tagged Image File Format, version 3</td>
<td>JPG&gt;TIF_2</td>
</tr>
<tr>
<td>JPG &gt; PNG</td>
<td>Portable Network Graphics, version 1.0</td>
<td>JPG&gt;PNG</td>
</tr>
<tr>
<td>JPG &gt; JPG_2000</td>
<td>JPG_2000</td>
<td>JPG&gt;JP2</td>
</tr>
</tbody>
</table>
Go/No-Go

- Deliberate step for taking a decision if it will be useful and cost-effective to continue the procedure, given
  - The resources to be spent (people, money)
  - The availability of tools and solutions,
  - The expected result(s).
- Review of the experiment/evaluation process design so far
  - Is the design complete, correct and optimal?
- Need to document the decision
- If insufficient: can it be redressed or not?
- Decision per alternative: go / no-go / deferred-go
Develop experiment

- Plan for each experiment
  - steps to build and test SW components
  - HW set-up
  - Procedures and preparation
  - Parameter settings, capturing measurements (time, logs...)

- Standardized Testbed-environment simplifies this step (PLANETS Testbed)

- Ideally directly accessible Preservation Action Services

- Ensures that results are comparable and repeatable
Run experiment

- Before running experiments: Test
- Call migration / emulation tools
- Local or service-based
- Capture process measurements (Start-up time, time per object, throughput, ...)
- Capture resulting objects, system logs, error messages,…
Develop and Run Experiment

**Result Files**

<table>
<thead>
<tr>
<th>Results</th>
<th>Sample Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIFF (tool A) ?</td>
<td>Polar bear 1</td>
</tr>
<tr>
<td></td>
<td>Polar bear 2</td>
</tr>
<tr>
<td>TIFF (tool B) ?</td>
<td>Polar bear 1</td>
</tr>
<tr>
<td></td>
<td>Polar bear 2</td>
</tr>
<tr>
<td>GIF (tool C) ?</td>
<td>Polar bear 1</td>
</tr>
<tr>
<td></td>
<td>Polar bear 2</td>
</tr>
<tr>
<td>PNG (tool D) ?</td>
<td>Polar bear 1</td>
</tr>
<tr>
<td></td>
<td>Polar bear 2</td>
</tr>
<tr>
<td>JPG &gt; BMP ?</td>
<td>Polar bear 1</td>
</tr>
<tr>
<td></td>
<td>Polar bear 2</td>
</tr>
<tr>
<td>JPG &gt; TIF ?</td>
<td>Polar bear 1</td>
</tr>
<tr>
<td></td>
<td>Polar bear 2</td>
</tr>
</tbody>
</table>

**Release 1.2 - Institute of Software Technology and Interactive Systems: « off-ice bears »**
Demo!
Evaluate experiment

- Analyse the results according to the criteria specified in the Objective Tree
- Preservation Characterization: Characterization Services
- Evaluation analyses
  - Experiment measurements, results
  - Necessity to repeat an experiment
  - Undesired / unexpected results
- Technical and intellectual aspects
In the first step of the workflow (Figure 8), the number of apps environment are described in the text boxes. The project (and organisation) can be altered on this page. Through
Figure 8 Define

In the first step of the workflow (Figure 8), the number of environment are described in the text boxes. The project
Evaluate Experiment

PLANETS Preservation Planning Planning Tool (*Plato*)

Evaluate Experiment
Expand All | Collapse All
Polar bear image preservation

<table>
<thead>
<tr>
<th>Focus</th>
<th>Node</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Polar bear image preservation</td>
</tr>
<tr>
<td>X</td>
<td>Process</td>
</tr>
<tr>
<td>X</td>
<td>Complexity</td>
</tr>
<tr>
<td>X</td>
<td>Cost</td>
</tr>
<tr>
<td>X</td>
<td>Image properties</td>
</tr>
<tr>
<td>X</td>
<td>Bits of colour depth</td>
</tr>
<tr>
<td>X</td>
<td>Technical characteristics</td>
</tr>
<tr>
<td>X</td>
<td>Official standard</td>
</tr>
<tr>
<td>X</td>
<td>Filesize (in Relation to Original)</td>
</tr>
</tbody>
</table>

Comments:

<table>
<thead>
<tr>
<th>Process &gt; Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative</td>
</tr>
<tr>
<td>TIFF (tool A)</td>
</tr>
<tr>
<td>TIFF (tool B)</td>
</tr>
</tbody>
</table>
| GIF (tool C) | Complex | *
| PNG (tool D) | Medium | *

<table>
<thead>
<tr>
<th>Process &gt; Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative</td>
</tr>
<tr>
<td>TIFF (tool A)</td>
</tr>
</tbody>
</table>
| TIFF (tool B) | 100 | *
| GIF (tool C) | 140 | *
| PNG (tool D) | 79 | *

Image properties > Bits of colour depth

| Alternative | Single result | Comments |
| TIFF (tool A) | 32 | bit |
| TIFF (tool B) | 32 | bit |
| GIF (tool C) | 32 | bit |
| PNG (tool D) | 32 | bit |
Practise time!

- Log into Plato at: http://www.ifs.tuwien.ac.at/dp/plato
- Download http://www.ifs.tuwien.ac.at/~kulovits/sample-files.zip
- Download http://www.ifs.tuwien.ac.at/~kulovits/experiment-results.zip
- Open preservation plan named “Scanned yearbooks archive (EVALUATE EXPERIMENTS)"
- Evaluate requirements
PP Workflow
Orientation

Preservation Planning in Plato

- Define requirements
  - Define basis
  - Choose records
  - Identify requirements
  - Tree templates and fragments

- Evaluate alternatives
  - Go/No-Go
  - Define alternatives
  - Develop experiment
  - Run experiment
  - Evaluate experiment
  - Mapping characteristics to requirements

- Consider results
  - Analyse results
  - Set importance factors
  - Transform measured values

- Preservation Action Recommendation

- Build preservation plan
  - Create executable preservation plan
  - Define preservation plan
  - Validate preservation plan

- Preservation Plan
Transform measured values

- Measures come in seconds, euro, bits, goodness values,…
- Need to make them comparable
- Transform measured values to uniform scale
- Transformation tables for each leaf criterion
- Linear transformation, logarithmic, special scale
- Scale 1-5 plus "not-acceptable"
Transform Measured Values

Polar bear image preservation

<table>
<thead>
<tr>
<th>Focus</th>
<th>Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>▼Polar bear image preservation</td>
<td></td>
</tr>
<tr>
<td>▼Process</td>
<td></td>
</tr>
<tr>
<td>▼Complexity</td>
<td></td>
</tr>
<tr>
<td>▼Cost</td>
<td></td>
</tr>
<tr>
<td>▼Image properties</td>
<td></td>
</tr>
<tr>
<td>▼Bits of colour depth</td>
<td></td>
</tr>
<tr>
<td>▼Technical characteristics</td>
<td></td>
</tr>
<tr>
<td>▼Official standard</td>
<td></td>
</tr>
<tr>
<td>▼Filesize (in Relation to Original)</td>
<td></td>
</tr>
</tbody>
</table>

Comments: The limit for the process cost was decided to be 150€ per picture in the last polar-bear-enthusiasts-association.

Technical characteristics > Official standard

<table>
<thead>
<tr>
<th>Ordinal Value</th>
<th>Target Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>5.0</td>
</tr>
<tr>
<td>No</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Aggregation mode:
- ◯ Worst result
- ◯ Arithmetic mean

Technical characteristics > Filesize (in Relation to Original)

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Target Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0</td>
<td>* x -&gt; 1</td>
</tr>
<tr>
<td>5.0</td>
<td>* x -&gt; 2</td>
</tr>
<tr>
<td>2.0</td>
<td>* x -&gt; 3</td>
</tr>
<tr>
<td>1.2</td>
<td>* x -&gt; 4</td>
</tr>
<tr>
<td>0.8</td>
<td>* x -&gt; 5</td>
</tr>
</tbody>
</table>

Results:
- TIFF (tool A): Yes
- TIFF (tool B): Yes
- GIF (tool C): No
- PNG (tool D): Yes
Orientation
Set Importance Factors

- Not all leaf criteria are equally important
- By default, weights are distributed equally
- Adjust relative importance of all siblings in a branch
- Weights are propagated down the tree to the leaves
Set Importance Factors

- Object characteristics
  - 35% menus and navigation path
    - 35% complete
    - 30% overall page layout
    - animated
    - pointer
    - speed
    - effects
    - graphic
    - sound
  - 25% structure
    - 25% mouse
    - 35% complete
    - 25% structure
  - 20% transitions
  - 15% animations
  - 5% background
  - 25% menu speed
  - type
  - colour
  - style
  - size
  - 10% appearance
  - 10% navigation
  - 80% reaction to activity
    - mouse
    - keyboard
  - 10% video/sound control
    - 10% video/sound control
  - 20% documentation material
  - 80% metadata reference valid
  - 10% context
    - Loops
    - Effects
    - content identical
    - quality
    - sound
    - 22% sound
      - 22% video
        - picture
        - synchronisation
        - resolution
        - colour
        - position
        - size
        - content identical
        - quality
        - 22% image
          - content identical
          - line breaks
          - 22% text
          - user manual
          - 12% user manual
          - 22% context
            - 22% content
              - 22% context
                - 22% content
                  - 22% sense
                    - 22% content
                      - 22% context
                        - 22% content
                          - 22% sense
                            - 22% content
                              - 22% context
                                - 22% content
                                  - 22% sense
                                    - 22% content
                                      - 22% context
                                        - 22% content
                                          - 22% sense
                                            - 22% content
                                              - 22% context
                                                - 22% content
                                                  - 22% sense
                                                    - 22% content
Analyse results

- Aggregate values in Objective Tree
  - Multiply transformed measurements in leaves with weights
  - Sum up across tree
- Results in accumulated performance value per alternative at root level
  → ranking of alternatives
- Also results in performance value for each alternative in each sub-branch of the tree
  → combination of alternatives
- Basis for well-informed and accountable decisions
- Different aggregation methods, e.g. sum and multiplication
### Analyse Results

**Planets Preservation Planning Tool (Plato)**

#### Analyse Results

**Aggregation method:** Sum

<table>
<thead>
<tr>
<th>Select</th>
<th>Alternative</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>PDF/A ToolA</td>
</tr>
<tr>
<td>✔</td>
<td>PDF/A ToolB</td>
</tr>
</tbody>
</table>

**Minimalist root node**

<table>
<thead>
<tr>
<th>Focus</th>
<th>Name</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>▼ Minimalist root node</td>
<td>PDF/A ToolA: 2,88, PDF/A ToolB: 3,19</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>Image properties</td>
<td>✔</td>
</tr>
<tr>
<td></td>
<td>PDF/A ToolA: 0,60, PDF/A ToolB: 0,80</td>
<td>✔️</td>
</tr>
<tr>
<td>X</td>
<td>Karma</td>
<td>✔️</td>
</tr>
<tr>
<td></td>
<td>PDF/A ToolA: 0,40, PDF/A ToolB: 0,00</td>
<td>✔️</td>
</tr>
<tr>
<td>X</td>
<td>Filesize (in Relation to Original)</td>
<td>✔️</td>
</tr>
<tr>
<td></td>
<td>PDF/A ToolA: 0,78, PDF/A ToolB: 0,99</td>
<td>✔️</td>
</tr>
<tr>
<td>X</td>
<td>A Single-Leaf</td>
<td>✔️</td>
</tr>
<tr>
<td></td>
<td>PDF/A ToolA: 0,40, PDF/A ToolB: 0,80</td>
<td>✔️</td>
</tr>
<tr>
<td>X</td>
<td>IntRange 0-10</td>
<td>✔️</td>
</tr>
<tr>
<td></td>
<td>PDF/A ToolA: 0,70, PDF/A ToolB: 0,60</td>
<td>✔️</td>
</tr>
</tbody>
</table>
**Analyse Results**

**PLANETS Preservation Planning Tool (Plato)**

**Aggregation method:** Multiplication

---

### Analyse Results

**How do the aggregation mechanisms work?**

<table>
<thead>
<tr>
<th>Select</th>
<th>Alternative</th>
</tr>
</thead>
<tbody>
<tr>
<td>✅</td>
<td>PDF/A ToolA</td>
</tr>
<tr>
<td>✅</td>
<td>PDF/A ToolB</td>
</tr>
</tbody>
</table>

**Minimalist root node**

<table>
<thead>
<tr>
<th>Focus</th>
<th>Name</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>▼Minimalist root node</td>
<td>PDF/A ToolA:2,86, PDF/A ToolB:0,00</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>▼Image properties</td>
<td>PDF/A ToolA:1,28, PDF/A ToolB:1,32</td>
</tr>
<tr>
<td>x</td>
<td>▼Amount of Pixel</td>
<td>PDF/A ToolA:3,50, PDF/A ToolB:4,00</td>
</tr>
<tr>
<td>x</td>
<td>▼Karma</td>
<td>PDF/A ToolA:1,15, PDF/A ToolB:0,00</td>
</tr>
<tr>
<td>x</td>
<td>▼Filesize (in Relation to Original)</td>
<td>PDF/A ToolA:1,31, PDF/A ToolB:1,38</td>
</tr>
<tr>
<td>x</td>
<td>▼A Single-Leaf</td>
<td>PDF/A ToolA:1,15, PDF/A ToolB:1,32</td>
</tr>
<tr>
<td>x</td>
<td>▼IntRange 0-10</td>
<td>PDF/A ToolA:1,28, PDF/A ToolB:1,25</td>
</tr>
</tbody>
</table>
Analyse results

Example: Electronic documents

<table>
<thead>
<tr>
<th>Alternative</th>
<th>Total Score Weighted Sum</th>
<th>Total Score Weighted Multiplication</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDF/A (Adobe Acrobat 7 prof.)</td>
<td>4.52</td>
<td>4.31</td>
</tr>
<tr>
<td>PDF (unchanged)</td>
<td>4.53</td>
<td>0.00</td>
</tr>
<tr>
<td>TIFF (Document Converter 4.1)</td>
<td>4.26</td>
<td>3.93</td>
</tr>
<tr>
<td>EPS (Adobe Acrobat 7 prof.)</td>
<td>4.22</td>
<td>3.99</td>
</tr>
<tr>
<td>JPEG 2000 (Adobe Acrobat 7 prof.)</td>
<td>4.17</td>
<td>3.77</td>
</tr>
<tr>
<td>RTF (Adobe Acrobat 7 prof.)</td>
<td>3.43</td>
<td>0.00</td>
</tr>
<tr>
<td>RTF (ConvertDoc 4.1)</td>
<td>3.38</td>
<td>0.00</td>
</tr>
<tr>
<td>TXT (Adobe Acrobat 7 prof.)</td>
<td>3.28</td>
<td>0.00</td>
</tr>
</tbody>
</table>

- Deactivation of scripting and security are knock-out criterium (PDF)
- RTF is weak in *Appearance* and *Structure*
- Plain text doesn’t satisfy several minimum requirements
PP Workflow
Practise time!

- Log into Plato at: http://www.ifs.tuwien.ac.at/dp/plato
- Open preservation plan named “Scanned yearbooks archive (ANALYSE)”
- Proceed to “Validate Preservation Plan”
- Export the preservation plan
Schedule

(1) Introduction
  - What is Digital Preservation?
  - EPrints
  - Preservation Planning and Plato

(2) Preservation in EPrints

(3) Preservation Planning with Plato

(4) Bringing it all together and Closing
**The Preservation Process**

- 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.

Preservation Plan Upload Successful

Actions have been queued to be executed shortly and changes will be reflected below once completed. In order to view these changes please revisit or refresh this page later.
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 Actions Panel

- Download plan for reviewing in planning software.
- Re-enact plan
Viewing the Result

- **Before**

  ![Image (GIF)](image-url)  
  Download (76Kb) | Preview

- **After**

  ![Image (PNG)](image-url)  
  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
Exercise Time
Conclusions
Preservation Planning

Why Preservation Planning?

- Several preservation strategies developed
  - For each strategy: several tools available
    - For each tool: several parameter settings available
- How do you know which one is most suitable?
- What are the needs of your users? Now? In the future?
- Which aspects of an object do you want to preserve?
- What are the requirements?
- How to prove in 10, 20, 50, 100 years, that the decision was correct / acceptable at the time it was made?
Preservation Planning

- Consistent workflow leading to a preservation plan
- Analyses, which solution to adopt
- Considers
  - preservation policies
  - legal obligations
  - organisational and technical constraints
  - user requirements and preservation goals
- Describes the
  - preservation context
  - evaluated preservation strategies
  - resulting decision including the reasoning
- Repeatable, solid evidence
Digital Preservation

What is a preservation plan?

- 10 Sections
  - Identification
  - Status
  - Description of Institutional Setting
  - Description of Collection
  - Requirements for Preservation
  - Evidence for Preservation Strategy
  - Cost
  - Trigger for Re-evaluation
  - Roles and Responsibilities
  - Preservation Action Plan

Preservation Plan Template
Preservation lifecycle

- Risk Analysis
- Preservation Planning
- Plan Enactment
- Re-Evaluation
Thank you!

http://www.ifs.tuwien.ac.at/dp

http://www.eprints.org/