The Preservation Planning Workflow:
From institutional requirements via experimental evidence to accountable preservation plans

Andreas Rauber
Department of Software Technology and Interactive Systems
Vienna University of Technology
http://www.ifs.tuwien.ac.at/~andi
Overview

- What is Digital Preservation?
- Preservation Planning
- Plato: Preservation Planning Tool
Why do we need Digital Preservation?
Why do we need Digital Preservation?
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.
Strategies for Digital Preservation

Strategies
(grouped according to Companion Document to UNESCO Charter
http://unesdoc.unesco.org/images/0013/001300/130071e.pdf)

- Investment strategies:
  - Standardization, Data extraction, Encapsulation, Format limitations

- Short-term approaches:
  - Museum, Backwards-compatibility, Version-migration, Reengineering

- Medium- / long-term approaches:
  - Migration, Viewer, Emulation

- Alternative approaches:
  - Non-digital Approaches, Data-Archeology

- No single optimal solution for all objects
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 (operating system, 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
What are some core DP challenges in scientific publishing wrt. the objects to be preserved?

- The publication
- The context of the publication
- Adjunct material (slides, notes, videos, …)
- Demos, exercises, interactive elements
- Data sets and simulations
- Community aspects, web 2.0, discussion forums, …
- Links to external material
- Other new functionalities
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 ages“
- 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
Overview

- What is Digital Preservation?
- Preservation Planning
- Plato: Preservation Planning Tool
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 Workflow

- Originally developed within the DELOS DP Cluster now refined and integrated within PLANETS
- Based on
  - Preservation Planning approach using utility analysis, developed at TU Vienna
  - Testbed for evaluation developed at Nationalarchief, The Netherlands
  - Follows the OAIS model
  - Consistent with requirements specified by ORLC/TRAC and Nestor criteria catalogue
Preservation Planning 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

Knowledge base
- Tree templates and fragments
- Mapping characteristics to requirements
- Preservation Plan

FACULTY OF INFORMATICS
Identify requirements

Analog…

… or

born
digital
Identify requirements
Identify requirements
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?)
Preservation Planning Workflow
Overview

- What is Digital Preservation?
- Preservation Planning
- Plato: Preservation Planning Tool
Preservation Planning with Plato

Plato

- Preservation Planning Tool
- Reference implementation of planning workflow
- Web-based application, release 2.0 Nov. 12 2008
- Documents the process and ensures that all steps are considered
- Automates several steps, integrates services
- Creates a preservation plan (XML, PDF)
- Technical basis:
  - Java Enterprise Beans, EJB 3 (Hibernate)
  - Based on JBoss Application Server
  - XML Import/Export (XStream)
- [http://www.ifs.tuwien.ac.at/dp/plato](http://www.ifs.tuwien.ac.at/dp/plato)
Define Sample Objects

Sample Records

Description of sample records: several samples of electronic theses

Sample Record

Full name: sample thesis 1
Short name: DA1
Has data: [download]
Original technical environment:
Description:

Object Format

PUID:
Name:
Version:
Mime-type:

Identify format

Add new record without file
Add record

Upload new record
Durchsuchen... Upload File

- Upload objects
- Identify using DROID & Pronom

FACULTY OF INFORMATICS
Create Objective Tree: Mindmaps & Tree-Editor

PLANEts Preservation Planning Tool (Plato)

Identify Requirements

Objective Tree
Descriptive Information

How can I define the objective tree?

Focus | Node | Single | Scale | Restriction | Unit
--- | --- | --- | --- | --- | ---
Website | Record characteristics | | | | |
Technical characteristics | | | | | |
Ubiquity | | | Ordinal | | |
Support | | | Positive Integer | | |
Documentation | | | | | |
Stability | | | | | |
Ease of identification | | | Ordinal | Automatic/Manual/No |
Ease of validation | | | Ordinal | Automatic/Manual/No |
Lossiness | | | Ordinal | Lossy/Lossless |
IPR | | | | | |
Complexity | | | Ordinal | None/Deterrable/Recoverable |

Release 1.1 - Institute of Software Technology and Interactive Systems, Vienna University of Technology
**Plato**

**PLANETS Preservation Planning Tool (Plato)**

### Define the alternatives of the Project

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Description</th>
<th>Remove</th>
</tr>
</thead>
<tbody>
<tr>
<td>196515</td>
<td>TIFF (tool A)</td>
<td>Convert to TIFF using the well-tested and expensive tool 'A'</td>
<td>Remove</td>
</tr>
<tr>
<td>196513</td>
<td>TIFF (tool B)</td>
<td>Convert to TIFF/4 using this new tool named 'B'</td>
<td>Remove</td>
</tr>
<tr>
<td>196514</td>
<td>GIF (tool C)</td>
<td>Convert to GIF using the well-tested tool 'C'</td>
<td>Remove</td>
</tr>
<tr>
<td>196515</td>
<td>PNG (tool D)</td>
<td>Convert to PNG using the well-tested tool 'D'</td>
<td>Remove</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; TIF</td>
<td>Tagged Image File Format, version 3</td>
<td>JPG&gt;BMP;TIF</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;TIF</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; JPEG</td>
<td>JPEG 2000</td>
<td>JPG&gt;JPEG</td>
</tr>
</tbody>
</table>

- **Select preservation actions**
- **Planets registry, Crib, Minimi**
- **Migration & Emulation services**
PLANEPEPS Preservation Planning Tool (Plato)

Analyse Results

<table>
<thead>
<tr>
<th>Focus</th>
<th>Name</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimalist root node</td>
<td>PDF/A (Tool A): 2.98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PDF/A (Tool B): 3.19</td>
<td></td>
</tr>
<tr>
<td>Image properties</td>
<td>PDF/A (Tool A): 0.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PDF/A (Tool B): 0.00</td>
<td></td>
</tr>
<tr>
<td>Vfirma</td>
<td>PDF/A (Tool A): 0.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PDF/A (Tool B): 0.00</td>
<td></td>
</tr>
<tr>
<td>Filesize (in Relation to Original)</td>
<td>PDF/A (Tool A): 0.76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PDF/A (Tool B): 0.06</td>
<td></td>
</tr>
<tr>
<td>Single-Leaf</td>
<td>PDF/A (Tool A): 0.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PDF/A (Tool B): 0.00</td>
<td></td>
</tr>
<tr>
<td>IntRange 0-10</td>
<td>PDF/A (Tool A): 0.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PDF/A (Tool B): 0.00</td>
<td></td>
</tr>
</tbody>
</table>

- Automatic evaluation of properties
- Analyze & compare performance
Deactivation of scripting and security are knock-out criterion (PDF)
- RTF is weak in Appearance and Structure
- Plain text doesn’t satisfy several minimum requirements
What we have after this presentation

- Basic Preservation Plan:
  - PDF: SummerSchool-Example - Final Report.pdf
  - XML: Summer_School_Preservation_Plan_for_Papers.xml

- That was developed in a solid, repeatable and documented process
- That is optimal for the needs of our institution and for the data at hand
Conclusions

- Preservation Planning to ensure “optimal” preservation
- A simple, methodologically sound model to specify and document requirements
- Repeatable and documented evaluation
- Basis for well-informed, accountable decisions
- Concretization of OAIS model
- Follows recommendations of TRAC and nestor
- Generic workflow that can easily be integrated in different institutional settings
- Plato:
  - Tool support to perform solid, well-documented analyses
  - Creates core preservation plan
  - 241 users who have created 135 preservation planning projects

http://www.ifs.tuwien.ac.at/dp
http://www.ifs.tuwien.ac.at/dp/plato
Tutorials / Training Events

- ECDL 2009, Sep. 27 – Oct. 2, Corfu, Greece
- Planets training, Sep. 21–23, Sofia, Bulgaria
- Planets training, June 22–24, Copenhagen, DK
- JCDL 2009, June 15–19, Austin/TX
- IST Africa 2009, May 6–8, Uganda
- DigCCurr 2009, April 1–3, Chapel Hill, USA
- wePreserve Training, March 23–27, Barcelona
- Nestor Spring School 2009, March 16–20, Stauffen, DE
- CeBIT 2009, Mar. 3–7, Hannover, Germany
- DPC repository day, Dec. 12, London
- ICADL 2008, Dec. 2–5, Bali, Indonesia
- RCDL 2008, Oct. 7–11, Dubna, Russia
- ECDL 2008, Sep. 14–19, Aarhus, Denmark
- DELOS Summer School 2008, June 8–11, Pisa, Italy
- Sun PASIG 2008, May 27–29, San Francisco, USA
Thank you!

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

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