Digital Preservation: How to Plan

Preservation Planning with Plato

Christoph Becker

Vienna University of Technology
http://www.ifs.tuwien.ac.at/~becker

Sofia, September 2009
Outline

• Why preservation planning?
  – Trusted digital repositories
  – Policies vs. plans

• Preservation Planning
  – What is a preservation plan?
  – How to create a preservation plan
  – The Planets Preservation Planning Workflow
  – Requirements definition
  – The planning tool Plato

• Part 2: Requirements discussion
Trustworthiness in digital repositories

- Consumers need trust in digital repositories
- Producers need trust in digital repositories
- Repositories need trust in external providers
- Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)
TRAC and Preservation Planning: Example

A 3.2 Repository has procedures and policies in place, and mechanisms for their review, update, and development as the repository grows and as technology and community practice evolve.

- Policies, plans, monitoring

A3.6 Repository has a documented history of the changes to its operations, procedures, software, and hardware that, where appropriate, is linked to relevant preservation strategies and describes potential effects on preserving digital content.

- Preservation plans need traceability
Definition of a Preservation Plan

• ‘A *preservation plan* defines a series of preservation actions to be taken by a responsible institution to address an identified risk for a given set of digital objects or records (called collection).’

• The Preservation Plan takes into account the preservation policies, legal obligations, organisational and technical constraints, user requirements and preservation goal. It also describes the preservation context, the evaluated alternative preservation strategies and the resulting decision for one strategy, including the rationale of the decision.
Objects in context

- Organisation
- Policy
- Preservation plan
- Preservation action
- Object
- Ownership
- Awareness
- Responsibility

Planets
Evaluating preservation strategies

- Variety of solutions and tools exist
- Each strategy has unique strengths and weaknesses
- Requirements vary across settings
- Decision on which solution to adopt is complex
- Documentation and accountability is essential

- Preservation planning assists in decision making
- Evaluating preservation strategies on representative samples according to specific requirements and criteria
Preservation Planning in Plato

- Web based planning tool implementing the Planets preservation planning workflow
- Publicly available
- Automation of the planning process
  - Integration of registries and services for
    - File format identification
    - Preservation action (migration, emulation...)
    - Characterisation and comparison
- Knowledge base to support planning
- Upcoming new release!
- [http://www ifs tuwien ac at/dp/plato](http://www.ifs.tuwien.ac.at/dp/plato)
Preservation planning environment

- Define requirements
  - Evaluate alternatives
    - Analyse results
      - Recommendation
        - Build preservation plan
          - Preservation plan

Knowledge base

- Monitor
  - requirements
  - technology
  - environment

- Objects
- Technology
- Usage criteria
- Policies
- Actions
- Repository
Define basis and samples

- Document basic assumptions and constraints
  - Mandate, objects, and designated community
  - Purpose of planning
  - Applying policies and constraints
  - Reasons for starting the planning process

- Collection
  - Size, type of objects, original environment, usage
  - Sample objects
Choose sample objects/records

- Define the set of objects that are the subject of preservation planning
  - Size of the collection
  - Growth rate
  - Object format
  - ...

- Specify representative sample objects that cover the variety of significant properties and technical characteristics
Define all relevant goals and characteristics (high-level, detail) for the situation

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, ...)

Put the objects in relation to each other (hierarchical)
- bottom-up
- top-down
Influence Factors

- Technology
- Standards
- User requirements
- Characteristics of digital objects

- Technical characteristics
  - Infrastructure characteristics
  - Process characteristics

- Requirements for preserving a collection of digital objects
  - Object characteristics
    - Content
    - Appearance
    - Structure
    - Behaviour
    - Context

- Legal constraints
- Policies
- Organisational requirements
- Business needs, Budget constraints
Stakeholders

- Input needed from a wide range of persons, depending on the institutional context and the collection.

- IT Staff
- Administration
- Managers
- Lawyers
- Domain experts
- Technical experts
- Consumers
- Others
- Producers
- Curators
- Technical characteristics
- Infrastructure characteristics
- Process characteristics
- Website
- Record characteristics
- Appearance
- Content
- Structure
- Behaviour
- Context

Logos:
- EU
- TU Vienna
- CEI
- Planets
An Objective Tree
Types of requirements

• Requirements on the outcome of actions
  – Access
  – Risks incurred
  – Format should be open, documented...
  – The objects should be
    • Authentic
    • Reliable
    • ...

• Requirements on the action
  – Fast
  – Reliable
  – Well supported
  – ...

[Image of diagram showing preservation planning environment, objects, technology, issues, materials, policies, actions, repository, and analysis results]
User perspective

• Goal of digital preservation is to serve (future) users in providing usable and authentic information

• What are needs/requirements of users?
  – easy access
  – knowledge about origin of documents/ to be able to interpret them
  – to use them for their own convenience

• Example requirements
  – some users prefer that all information is presented in a uniform way
  – some users prefer that they can search full-text in documents (consequence: don’t migrate texts to image files)
  – ...

EU | ! | TU VIENNA | CEI | Planets
Requirements for objects

- Authenticity
- Reliability
- Integrity
- Usability
- Accuracy
Essential characteristics of ‘digital objects’

• What needs to be preserved?
  – Content
  – Context
  – Structure
  – Appearance
  – Behaviour
Assign Measurable Units

- Leaf criteria should be objectively measurable
  - Seconds per object
  - Euro per object
  - Bits of colour depth

- Subjective scales where necessary
  - Adoption of file format
  - Amount of (expected) support

➤ Quantitative results
## Objective Tree

**PLANETS Preservation Planning Tool (Plato)**
Institute of Software Technology and Interactive Systems

### Identify Requirements

<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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Appearance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Content</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Behaviour</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Deactivate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Preserve</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Menus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Pop-ups</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Website</td>
<td>▼Freeze</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Current date/time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Visitor counter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Newsfeeds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Content</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Technical characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Usability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Tool Support</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Documentation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Stability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Ease of identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>▼Ease of validation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© 2007 Institute of Software Technology and Interactive Systems; «office bears»
Evaluate alternatives

- List applicable actions
  - Migration
  - Emulation
  - Both, other...

- Develop and run an experiment
  - Apply each action to each sample
  - Measure effects
  - Evaluate outcome
Discovering possible actions

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&gt;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_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; JP2</td>
<td>JPEG 2000</td>
<td>JPG&gt;JP2</td>
</tr>
</tbody>
</table>
Develop and run experiment

• Formulate for each experiment detailed
  – procedures and preparation
  – parameter settings for integrating preservation services
  – Evaluation/experiment plan (workflow/sequence of activities)

• Apply the selected potential preservation actions on the sample objects
  – Partly automated by web services
  – Partly manual
Evaluate experiment

- Evaluate the outcome of each alternative for each leaf of the objective tree
- Partly automated by tool support
  - Comparing objects: XCL, Jhove, ImageMagick, ...
  - Measuring performance
  - Judging file formats
  - ...
- Result: evaluated tree
Preservation planning environment

- Define requirements
- Evaluate alternatives
- Analyse results
- Recommendation
- Build preservation plan
- Preservation plan

Knowledge base
- Monitor
  - requirements
  - technology
  - environment

Objects
Technology
Usage criteria
Policies
Actions
Repository
Transform measured values

- Measures come in seconds, euro, bits, ...
- Need to make them comparable
- Transform measured values to uniform scale
- Target scale 0-5
- Two types of transformation
  - Numeric
  - Ordinal
- Result: tree ready for analysis
Set importance factors

PLANETS Preservation Planning Planning Tool (Plato)

Set Importance Factors
Balance weights automatically

Object characteristics

<table>
<thead>
<tr>
<th>Focus</th>
<th>Name</th>
<th>Weight</th>
<th>Lock</th>
<th>Total weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>▼</td>
<td>Object characteristics</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>X</td>
<td>behaviour</td>
<td>0</td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>X</td>
<td>structure</td>
<td>0</td>
<td></td>
<td>0.25</td>
</tr>
<tr>
<td>X</td>
<td>context</td>
<td>0</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>X</td>
<td>appearance</td>
<td>0</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>X</td>
<td>content</td>
<td>0</td>
<td></td>
<td>0.4</td>
</tr>
</tbody>
</table>

Save Proceed
Analyse Results

- Aggregate values
  - Multiply the transformed measured values in the leaf nodes with the leaf weights
  - Sum up the transformed weighted values over all branches of the tree
- Rank alternatives according to overall performance value at root
- Performance of each alternative
  - overall
  - for each sub-criterion (branch)
- Comparison of different alternatives
## Results: Weighted sum

**Result-Tree with all Alternatives, Aggregation method: Weighted sum.**

The tree contains only strategies that do not have knock-out evaluation criteria; see above.

### Analyse results

<table>
<thead>
<tr>
<th>Focus</th>
<th>Name</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polar bear image preservation</td>
<td>TIFF (tool A): 4,78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool B): 4,28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PNG (tool D): 3,97</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool A): 4,75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool B): 4,16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PNG (tool D): 0,74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool A): 3,50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool B): 2,50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PNG (tool D): 1,25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool A): 5,00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool B): 5,00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PNG (tool D): 5,00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool A): 1,43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool B): 1,43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PNG (tool D): 1,43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool A): 3,50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool B): 3,50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PNG (tool D): 3,50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool A): 0,83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIFF (tool B): 0,83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PNG (tool D): 1,12</td>
<td></td>
</tr>
</tbody>
</table>

### Conclusion

**Recommendation**

- **Recommendation:**
  - **Reasoning:**
  - **Effects of applying this strategy:**
Questions?

becker@ifs.tuwien.ac.at

www.ifstuwien.ac.at/dp/plato
www.planets-project.eu
Scenario: We need a plan

- The purpose of planning is to find a strategy on how to preserve a collection for the future, i.e. choose a tool to handle our collection with.

- The tool must be compatible with our existing hardware and software infrastructure, to install it within our server and network environment.

- The files haven't been touched for several years now and no detailed description exists. However, we have to ensure their accessibility for the next years.

- ‘A preservation plan defines a series of preservation actions to be taken by a responsible institution to address an identified risk for a given set of digital objects or records (called collection).‘
Scenario: Scanned images

- Discussion scenario for today: Scanned images
- Specific exercise scenario tomorrow: Create a preservation plan for a collection of scanned images
- General characteristics of this scenario
- Mission statement
- High-level requirements
High level requirements 1/2

- **Formats**
  - must/shall be standardised...
  - Compression?
- **Tools**
  - must/shall be open source,
  - Must not cost more than...
- **Bit-stream preservation costs...**
  - Depend on the file size and other factors
  - Must not exceed ... (per object)
- **Strategy**
  - consider migration ,
  - consider emulation (copyright?)
High level requirements 2/2

• Objects must be
  – “the same” – “unchanged” – “authentic” …
  – Significant properties need to be defined and measured
  – Content, context, structure, appearance, behaviour

• Trade-offs might be necessary
  – Usability vs. authenticity
  – Structure vs. independency
  – Access vs. costs
  – …
Analog...
... or born-digital
Requirements for objects

- **Authenticity**
  - to be what it purports to be,
  - to have been created or sent by the person purported to have created or sent it, and
  - to have been created or sent at the time purported
- **Reliability**
  - contents can be trusted as a full and accurate representation of the transactions, activities or facts to which they attest and can be depended upon in the course of subsequent transactions or activities
- **Integrity**
  - being complete and unaltered
- **Usability**
  - can be located, retrieved, presented and interpreted
- **Accuracy**
  - the degree to which data, information, documents or records are precise, correct, truthful, free of error or distortion or pertinent to the matter.
Essential characteristics of ‘digital objects’

- What needs to be preserved?
  - Content
  - Context
  - Structure
  - Appearance
  - Behaviour
Questions?

becker@ifs.tuwien.ac.at

www ifs tuwien ac at/dp/plato

www planets project eu
Usage

- Access
  - Main content
    - Searchable: Boolean
    - Indexable: Boolean
  - Metadata
    - Searchable: Boolean
    - Indexable: Boolean
  - Content
    - Changeable: Boolean
    - Annotate: Boolean
  - Flexibility
  - Manageability
    - Physical availability: Boolean
    - Time required to access: Integer (seconds or milliseconds)
  - Technical
    - File size: Larger/Same/Smaller
  - Provenance
    - Design and look: Same/Different
  - Metadata
    - Completeness: Boolean
    - Accuracy: Boolean
  - Changes documented: Boolean

PP/3 - User requirements
Static web pages

Website

Technical characteristics
- Ubiquity
- Support
- Documentation
- Stability
- Ease of identification
- Ease of validation
- Lossiness
- IPR
- Complexity
- Error tolerance
- Comparative size
- Acquisition
- Maintenance
- Hardware
- Software

Infrastructure characteristics
- Training
- Operations
- Staff
- Automation
- Streamable
- Throughput
- Performance
- Scalability
- Format
- Content
- Usability
- Syntactic
- Semantic

Process characteristics

Record characteristics
- Appearance
- Page layout
- Links
- Highlighting
- Text
- Cells
- Documents
- Text content
- Encoding
- Images
- Width
- Height
- Resolution
- Bit depth
- Colour space
- Histogram
- Internal links present and working
- Rows unchanged
- Columns unchanged
- Cell content unchanged
- Tables
- Pagination
- No. of pages
- Page breaks
- Structure
- Document structure
- Headers
- Footers
- TOC
- Appendices
- Sequence
- Behaviour
- Deadcode
- Mailto:
- Present
- Menus
- Pop-ups
- Current date/time
- Visitor counter
- Newsfeeds
- External links present and unchanged
- Links within the collection present and valid
• Visitor counter and similar things can be
  • Frozen at the point of harvesting
  • Left out
  • Still counting while being accessed in the archive
    (Is this desirable?)
Interactive multimedia

35% menus and navigation path
  Y/N 35% complete
  Y/A/N 30% overall page layout

Object characteristics

15% behaviour
  10% navigation
  80% reaction to activity
  10% video/sound control

10% context
  20% documentation material
  80% metadata reference valid

40% content
  22% sound
    Loops
    Effects
    content identical
    quality
    sound
  22% video
    picture
    synchronisation
  22% image
  22% text
  12% user manual

25% mouse
20% transitions
15% animations
5% background
25% menu speed
10% fonts
10% appearance

35% complete

25% structure
### Behaviour

- Interactive presentations exhibit two facets
  - Graph-like navigation structure
  - Navigation along the paths

<table>
<thead>
<tr>
<th>Node</th>
<th>Scale</th>
<th>Restriction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>behaviour</td>
<td></td>
<td></td>
</tr>
<tr>
<td>navigation</td>
<td>Ordinal</td>
<td>interactive and integrated/navigatable/none</td>
</tr>
<tr>
<td>reaction to activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mouse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>position</td>
<td>Boolean</td>
<td>✓</td>
</tr>
<tr>
<td>clicks</td>
<td>Boolean</td>
<td>✓</td>
</tr>
<tr>
<td>keyboard</td>
<td>Boolean</td>
<td>✓</td>
</tr>
<tr>
<td>video/sound control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>menus and navigation path</td>
<td>Ordinal</td>
<td>complete and free/partial (linear)/none</td>
</tr>
<tr>
<td>complete</td>
<td>Boolean</td>
<td>✓</td>
</tr>
<tr>
<td>overall page layout</td>
<td>Ordinal</td>
<td>Y/A/N</td>
</tr>
</tbody>
</table>
The content of a preservation plan

1. Identification
2. Status
   ✓ What was the immediate reason for this plan?
   ✓ Has it been approved and if so, when and by whom
   ✓ How does it relate to other P-plans related to a specific type of objects?
3. Description of institutional setting
4. Description of the collection (digital objects)
5. Purpose and requirements
6. Evidence of decision for a specific preservation action
   ✓ what is the foundation of the decision
   ✓ description of evaluation of possible actions
7. Costs considerations
8. Trigger for re-evaluation
9. Roles and responsibilities
10. Preservation action plan
    ✓ executable program