The Planets Preservation Planning workflow and the planning tool Plato

March, 27th 2009
Hannes Kulovits
Vienna University of Technology
http://www.ifs.tuwien.ac.at/~kulovits
Outline

• Preservation Planning
  – Evaluation of potential actions

• The Planets Preservation Planning Workflow
  – Workflow walkthrough
  – Requirements definition
  – The planning tool Plato
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
Scenario: Changes in user community

- Repository of electronic publications
- Policy: 90% of users can access all published reports
- Usage profile: 98% of users cannot view \texttt{dvi} files
- Content profile: 5% of published reports in \texttt{dvi} format

- Mission: Build and execute a plan for preserving access to these documents for the designated user community
Planets
Preservation Planning Workflow

- Define requirements
- Evaluate potential actions
- Analyse results
- Build a preservation plan
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
Define basis

- Document basic assumptions and constraints
  - Types of objects
  - Purpose of planning
  - Mandates and designated community
  - Applying policies
  - Triggers that initiated the planning process
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
Identify requirements

- Define all relevant goals and characteristics (high-level, detail) with respect to a 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, ...)

- Put the objects in relation to each other (hierarchical)

- Objective tree approaches:
  - bottom-up
  - top-down
Influence Factors

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

Requirements for preserving a collection of digital objects

- Technical characteristics
  - Infrastructure characteristics
  - Process characteristics
- 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
Domain experts
Curators
IT Staff
Domain experts
Curators
IT Staff
Domain experts
Curators
IT Staff
Domain experts
Curators
IT Staff
Domain experts
Curators
IT Staff
Domain experts
Curators
... or born-digital
Case Study: Web archiving

- Static web pages from the public domain
- Includes documents in formats such as doc, pdf
- Images
- No interactive content shall be preserved
A bit more detail…
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
Types of scales

- Numeric (unit)
- Yes/No (Y/N)
- Yes/Acceptable/No (Y/A/N)
- Ordinal: define the possible values (good/bad/ugly)
- Subjective 0-to-5 (0/5)
- **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?)
**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>
<tr>
<td>Focus</td>
<td>Node</td>
<td>Single</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Website</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Record characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appearance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behaviour</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inactivate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Email</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preserve</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Menu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pop-ups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freeze</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current date/time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visitor counter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newsfeeds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Context</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ubiquity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tool Support</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Documentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ease of identification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ease of validation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The image contains a web interface for a preservation planning tool, with specific focus areas and nodes. The table lists various elements with their respective options for scale and restriction, such as Boolean, Ordinal, and Positive Number. The image is from the Institute of Software Technology and Interactive Systems.
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

Tree templates and fragments
Mapping characteristics to requirements
Knowledge base
Define alternatives

- Given the type of objects and requirements, what strategies would be best suitable/are possible?
  - Migration
  - Emulation
  - Both
  - Other?

- For each alternative precise definition of
  - Which tool (OS, version,...)
  - Which functions of the tool in which order
  - Which parameters
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>

Create alternatives for selected services
Specify resources

- Detailed design and overview of the resources for each alternative
  - human resources (qualification, roles, responsibility, …)
  - technical requirements (hardware and software components)
  - time (time to set-up, run experiment, …)
  - cost (costs of the experiments, …)
Go/No-Go

• Deliberate step for taking a decision whether 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 readdressed or not?
Develop and run experiment

• Formulate for each experiment detailed
  • Development plan
    • steps to build and test software components
    • 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
Evaluate experiment

• Evaluate the outcome of each alternative for each leaf of the objective tree

• The evaluation will identify
  • Need for repeating the process
  • Unexpected (or undesired) results

• Includes both technical and intellectual aspects

• Evaluation may include comparing the results of more than one experiment/evaluation.
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
Transform measured values

- Measures come in seconds, euro, bits, goodness values, …
- Need to make them comparable
- Transform measured values to uniform scale
- Target scale 0-5
## 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>▼ Object characteristics</td>
<td>0</td>
<td><img src="image" alt="1" /></td>
<td><img src="image" alt="×" /></td>
<td><img src="image" alt="1" /></td>
</tr>
<tr>
<td><img src="image" alt="×" /></td>
<td>behaviour</td>
<td>0</td>
<td><img src="image" alt="×" /></td>
<td><img src="image" alt="0.15" /></td>
</tr>
<tr>
<td><img src="image" alt="×" /></td>
<td>structure</td>
<td>0</td>
<td><img src="image" alt="×" /></td>
<td><img src="image" alt="0.25" /></td>
</tr>
<tr>
<td><img src="image" alt="×" /></td>
<td>context</td>
<td>0</td>
<td><img src="image" alt="×" /></td>
<td><img src="image" alt="0.1" /></td>
</tr>
<tr>
<td><img src="image" alt="×" /></td>
<td>appearance</td>
<td>0</td>
<td><img src="image" alt="×" /></td>
<td><img src="image" alt="0.1" /></td>
</tr>
<tr>
<td><img src="image" alt="×" /></td>
<td>content</td>
<td>0</td>
<td><img src="image" alt="×" /></td>
<td><img src="image" alt="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.
This tree contains only strategies that do not have knock-out evaluation criteria; see above

### Polar Bear Image Preservation

<table>
<thead>
<tr>
<th>Focus</th>
<th>Name</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td></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): 3,87</td>
<td></td>
</tr>
</tbody>
</table>

### Process

|                  | TIFF (tool A): 1,03 |                         |
|                  | TIFF (tool B): 1,16  |                         |
|                  | PNG (tool D): 0,74   |                         |

### Complexity

|                          | TIFF (tool A): 2,50  |                         |
|                          | TIFF (tool B): 2,50  |                         |
|                          | PNG (tool D): 1,25   |                         |

### Cost

|                  | TIFF (tool A): 2,50  |                         |
|                  | TIFF (tool B): 1,00  |                         |
|                  | PNG (tool D): 1,00   |                         |

### Image Properties

|                   | TIFF (tool A): 1,70  |                         |
|                   | TIFF (tool B): 1,70  |                         |
|                   | PNG (tool D): 1,70   |                         |

### Bits of Colour Depth

|                  | TIFF (tool A): 5,00  |                         |
|                  | TIFF (tool B): 5,00  |                         |
|                  | PNG (tool D): 5,00   |                         |

### Technical Characteristics

|                  | TIFF (tool A): 1,43  |                         |
|                  | TIFF (tool B): 1,43  |                         |
|                  | PNG (tool D): 1,53   |                         |

### Official Standard

|                  | TIFF (tool A): 3,50  |                         |
|                  | TIFF (tool B): 3,50  |                         |
|                  | PNG (tool D): 3,50   |                         |

### Filesize (in Relation to Original)

|                  | TIFF (tool A): 0,83  |                         |
|                  | TIFF (tool B): 0,83  |                         |
|                  | PNG (tool D): 1,12   |                         |

## Conclusion

### Recommendation

| Recommendation: |                         |
| Reasoning:      |                         |

### Effects of applying this strategy:

|                         |                         |
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

Knowledge base

Tree templates and fragments
Mapping characteristics to requirements

Preservation Plan
Create executable plan

- Preservation Action Plan
- When?
  - Conditions and triggers for execution
  - Hardware and software requirements…
- What?
  - Single tool, composite workflow of services…
  - Validation and QA
  - Other actions needed, such as reporting…
Define preservation plan

• Executable action plan is not enough
  • Rules for monitoring
  • Evidence of decisions
  • Estimates of costs
  • Roles and responsibilities
  • …
Validate preservation plan

- Validate all elements of the plan
- Check for completeness
- Formally approve the plan and put it into action

- Continuous review and monitoring is necessary!
Summary

Preservation planning environment

- Objects
- Technology
- Usage criteria
- Policies
- Actions

Define requirements

Evaluate alternatives

Analyse results

Recommendation

Build preservation plan

Preservation plan

Knowledge base

Monitor
- requirements
- technology
- environment

Repository
Plato is openly available at:

http://www.if.s.tuwien.ac.at/dp/plato
Thank you

kulovits@ifs.tuwien.ac.at

www.ifs.tuwien.ac.at/dp/plato
www.planets-project.eu