Preservation planning 1
What to decide and how

April 8, 2013
Christoph Becker
Vienna University of Technology
www ifs tuwien ac at/~becker
Digital preservation and planning

- The mission of digital preservation
  - keep content authentic and understandable for a user community over time

- The mission of preservation planning
  - What to do and how
  - Defining the right courses of actions
  - Questions:
    • How to select the right action in a given scenario?
    • How to ensure trust?
    • How to enable scalability?
    • How to ensure continuous alignment over time?
Preservation Planning 1: Agenda

- The Problem: What to decide and how?
  - Decision problems in digital preservation
- The context: Goals, drivers, constraints
- From goals to actions: Preservation planning overview
  - What is preservation planning?
  - What is a preservation plan?
  - How to create a preservation plan, part 1
- DPÜ Exercise Task 2
Preservation Operations

- Core operations for preservation
  - Analyse content
  - Perform preservation actions
  - Perform Quality Assurance
  - Manage metadata
  - Report
Evaluating preservation actions

- Several actions available (migration, emulation, ...)
- Challenges:
  - Quality varies across tools
  - Properties vary across content
  - Usage varies across communities
  - Requirements vary across scenarios
  - Risk tolerance varies across collections
  - Preferences and constraints vary across organisations
  - Cost structures and compatibility varies across environments
  - Constraints, priorities and requirements shift constantly

- Systematic software component evaluation
How can we select the optimal preservation action for a given setting?

- What are the drivers and constraints on the decision space?
- What are the goals and objectives?
- What are the factors influencing the decision makers’ preferences?
- How can we model multiple competing objectives and requirements?
- How should we evaluate software components?
How can we ensure trustworthy preservation planning?

- What are the requirements on trust that need to be addressed?
- What decision steps and evidence need to be documented?
- What are the aspects that a plan needs to address, and what are the elements needed to cover them?
- How can we ensure reliable evaluation procedures and repeatable evidence?
How can we ensure that decision processes scale up?

- How can we automate decision making?
- How can we integrate continuous monitoring?
- Which properties can be measured automatically, and how?
- How can we create a controlled environment for observing the behaviour of components in a reproducible way?
Evaluating preservation actions

- **Context: Trustworthy repositories**
  - Open Archival Information Systems model (OAIS)
  - Trustworthy repositories criteria (TRAC, nestor)
  - Trust requires evidence
  - Evidence needs repeatable, objective facts

- **Preservation planning approach**
  - Evaluate potential actions objectively against scenario-specific requirements in a repeatable way
  - Sample-based experiments in controlled environment
  - Quantitative analysis of strengths and weaknesses
  - Evaluate suitability of each potential action
Challenges in Component based software development:

- Importance of non-functional requirements
- Dynamism
- Importance of trust
- Functional granularity

--- CBSD
Web service selection

Challenges
Preservation planning

Importance of non-functional requirements

Correctness!

Dynamism

Functional granularity

Importance of trust

- CBSD
- WS
- Plato
Operations in context

- Preservation Context and Management
- Preservation Monitoring
- Preservation Planning
- Preservation Operations

- Policies
- Content Profiling c3po
- Plato

- Why? How?
- Plan
- Migrate?
- emulate?
- manual...
- ... automated

- Workflow automation, Taverna, ....
Trustworthy preservation planning

- Preservation planning as a capability:
  - the ability to assess the impact of influencers and specify actionable preservation plans that define concrete courses of actions and the directives governing their execution
  - the operative management of obsolescence to maximize expected value with minimal costs

- A preservation plan specifies actions
  - scope and what, how, when, who, why

- Trust requires evidence
  - Trust has to be evaluated in a realistic context
    - Documented evidence
    - Controlled experimentation
    - Scenario-specific requirements assessment
Preservation planning environment

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

Knowledge base
- Monitor
  - requirements
  - technology
  - environment

- Objects
- Technology
- Usage criteria
- Policies
- Actions

Repository
Preservation Planning: Key concepts

- Repeatable, standardized planning workflow
- A weighted hierarchy of objectives
  - Measurable criteria on the leaf level of the tree
  - Utility functions make criteria comparable
- Controlled experimentation on sample content
  - Evidence-based decision making
- Standardized structure for plan specification
  - Transparency and documentation
  - Comparability across scenarios
  - Integration with repository systems
- Planning tool Plato guides, validates, documents
  - www.ifs.tuwien.ac.at/dp/plato
- Automation: Reduce manual effort
What is 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 goals.

- It also describes the preservation context, the evaluated alternative preservation strategies and the resulting decision for one strategy, including the rationale of the decision.
Levels of control

- **Guidance**
  - Strategies
  - Compliance
  - Regulations

- **Control**
  - Specific preservation policies

- **Planning**
  - Preservation Planning
  - Monitoring

- **Operations**
  - Preservation actions
  - Quality Assurance
  - Content analysis
Example scenarios for planning

- Digital photography archive with diverse camera RAW files from many cameras
  - Should we normalise to reduce risk and facilitate access? To DNG? To TIFF? How?

- Console video games
  - Use emulation? Which emulator and how?

- Large-scale digitized newspaper archive in non-standard TIFF
  - Convert to JPEG2000 or TIFF6 compressed to reduce costs? Leave unchanged and wait?

- Interactive digital art
  - What to do?
# Digital Preservation drivers

Table 4 - DP drivers as specified in the SHAMAN Reference Architecture

<table>
<thead>
<tr>
<th>Internal</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Vision</td>
<td>Goals, Scope of designated community, etc.</td>
</tr>
<tr>
<td>Resources</td>
<td>Infrastructure (e.g., operational costs, expertise needed), Hardware (e.g., operational costs, technological capability), Software (e.g., operational costs, technological capability), Staff (e.g., expertise and qualifications, commitment)</td>
</tr>
<tr>
<td>Data</td>
<td>Volume, Structure, Representation, Semantics, etc.</td>
</tr>
<tr>
<td>Processes</td>
<td>Dependencies, Responsibilities, Alignment, etc.</td>
</tr>
<tr>
<td>Producers</td>
<td>Demand satisfactions, Content, Technology, Trust and reputation</td>
</tr>
<tr>
<td>User community</td>
<td>Technology, Knowledge, Demand satisfaction, Trust and reputation</td>
</tr>
<tr>
<td>Contracts</td>
<td>Deposit, Supplier and service, Interoperability, Access, etc.</td>
</tr>
<tr>
<td>Supply</td>
<td>Technology, Services, People</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>External</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Competition</td>
<td>Overlap of: Services, Content, User community, Producers, Technology, Mandate, Rights, Funding, Capabilities</td>
</tr>
<tr>
<td>Regulation and mandate</td>
<td>Regulation/Legal constraints, Embedding organization regulation, Mandate, Rights and ownership, Certification, Funding</td>
</tr>
</tbody>
</table>
A scenario in a National Library

Scanned old print material, digitally recorded AV broadcast material, web archive
Legal mandate for long-term preservation and access, IT department in-house

Drivers and constraints?

<table>
<thead>
<tr>
<th>Internal</th>
<th>Business Vision</th>
<th>Goals, Scope of designated community, etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resources</td>
<td>Infrastructure (e.g., operational costs, expertise needed), Hardware (e.g., operational costs, technological capability), Software (e.g., operational costs, technological capability), Staff (e.g., expertise and qualifications, commitment)</td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>Volume, Structure, Representation, Semantics, etc.</td>
<td></td>
</tr>
<tr>
<td>Processes</td>
<td>Dependencies, Responsibilities, Alignment, etc.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>External</th>
<th>Producers</th>
<th>Demand satisfactions, Content, Technology, Trust and reputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>User community</td>
<td>Technology, Knowledge, Demand satisfaction, Trust and reputation</td>
<td></td>
</tr>
<tr>
<td>Contracts</td>
<td>Deposit, Supplier and service, Interoperability, Access, etc.</td>
<td></td>
</tr>
<tr>
<td>Supply</td>
<td>Technology, Services, People</td>
<td></td>
</tr>
<tr>
<td>Competition</td>
<td>Overlap of: Services, Content, User community, Producers, Technology, Mandate, Rights, Funding, Capabilities</td>
<td></td>
</tr>
<tr>
<td>Regulation and mandate</td>
<td>Regulation/Legal constraints, Embedding organization regulation, Mandate, Rights and ownership, Certification, Funding</td>
<td></td>
</tr>
</tbody>
</table>
Preservation policy: examples

- Policies are elements of governance
- They “govern”, i.e. **guide, shape and control** preservation operations

- Example policy statements of institutions with a digital preservation programme
  - UK Data Archive
  - National Archives of Australia
  - ISO/TR 18492:2005
    Long-term preservation of electronic document-based information
UK Data Archive

- UK Data Archive Preservation Policy
  - p. 11: “The UKDA has chosen to implement a preservation strategy based upon open and available file formats, data migration and media refreshment.”

- What does this choice mean in practice?
- Two examples:
  - Emulation is –apparently– not a preservation strategy that will be chosen; all obsolete files will be migrated.
  - Migration to open file formats will be preferred.
An Approach to the Preservation of Digital Records

p. 14: “The digital preservation program must be able to preserve any digital record that is brought into National Archives’ custody regardless of the application or system it is from or data format it is stored in.”

What does this choice mean in practice?

For example:
  – all records that are accepted should be preserved, regardless of file format, medium, application, etc.
  – transform to open standard + keep ‘original’ format
International standard: Long-term preservation of electronic document-based information

p. 12: Migration to standard formats

Storage repositories should consider migrating electronic document-based information from the wide variety of formats used by creators or recipients to a smaller number of “standardized” formats upon their transfer to the custody of the repository.

“Standardized” formats could be a consensus on formats that are widely used and are likely to cover a majority of a particular class of electronic document-based information. Proprietary file formats should be avoided. Among the technology neutral formats that merit consideration are PDF/A-1, XML, TIFF and JPEG.
From guidance to action (*simplified*)

- **Guidance**
  - Open formats
  - Migration

- **Control**
  - Formats must be ISO-standardised
  - No lossy compression may be used

- **Planning**
  - Concrete criteria and decision

- **Operations**
  - Migrate to PDF/A
  - Quality assure results
What is *in* a preservation plan?

- Definition of scope and context
  - What to preserve and why
- Definition of objectives
  - What to achieve
- Set of actions, evaluation and recommendation
  - How to preserve it and why
- Documentation of actions and reasons
  - Why did we decide what
- Conditions for QA and monitoring
  - What to look out for
Characteristics of a preservation plan

- Translation of a preservation policy
- Specification of how to treat a collection in a given setting
- Monitored for
  - changes in technology
  - changes in organisational setting
  - changes in user requirements
  - changes in available tools
  - changes in preservation methods
- Species concrete action
  - The **preservation action plan** can be an executable workflow definition, detailing actions and required technical environment
  - The preservation plan provides the context/background of the preservation action plan
Agenda

- The Problem: What to decide and how?
  - Decision problems in digital preservation
- Preservation planning overview
  - What is preservation planning?
  - What is a preservation plan?
  - How to create a preservation plan: Part 1
- DPÜ Exercise Task 2
Preservation planning environment

- Objects
- Technology
- Usage criteria
- Policies
- Actions

1. Define requirements
2. Evaluate alternatives
3. Analyse results
4. Recommendation
5. Build preservation plan

Knowledge base
- requirements
- technology
- environment

Monitor

Repository

Preservation plan
Define basis

- What are the objects?

- What are the fundamental requirements?
  - Authenticity, reliability, integrity, usability
  - Metadata (for different purposes)

- What are the applying policies, legal constraints, regulations...
  - User groups, target community
  - Institutional settings
Define sample objects

- Representative for the objects in the collection
- 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

... c3po
Select content type
• e.g.: Legal documents from the enterprise archive

Select properties
• Property set determined by type “documents” (page count, …)

High-level issue detection
• Object-level policy violations: Validity, encryption, …
• Collection-level: format normalisation…

Select scoping property
• Subformat: PDF 1.2…
• Other properties: all protected documents…. 

Select samples
• Single dimension: page count, size, age, validity…
• Multiple dimensions: Largest invalid, oldest protected …
An example collection of PDF documents

Collection size: ~40,000 files

- 2 files "Unknown Binary"
- 1 file exposed as MS Word DOC
- 1 file with size 0
Creation date per PDF file type
Which two files are similar?

- Consider three files A, B, C

<table>
<thead>
<tr>
<th>Format</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PDF 1.2</td>
<td>PDF 1.2</td>
<td>PDF 1.4</td>
</tr>
</tbody>
</table>
Which two files are similar?

- Consider three files A, B, C

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>PDF 1.2</td>
<td>PDF 1.2</td>
<td>PDF 1.4</td>
</tr>
<tr>
<td>Page count</td>
<td>20</td>
<td>20,000</td>
<td>40</td>
</tr>
<tr>
<td>Encryption</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>File size</td>
<td>1MB</td>
<td>120 MB</td>
<td>2 MB</td>
</tr>
<tr>
<td>Valid</td>
<td>no</td>
<td>yes</td>
<td>No</td>
</tr>
<tr>
<td>Well-formed</td>
<td>Yes</td>
<td>yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Digital signature</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

- … file format is just one property
- Careful selection is critical
Requirements definition

- What are our goals and objectives?
- How do we measure achievement of our goals?
- Which drivers have an influence on which objectives?
- Define complementary criteria for all objectives
  - Trade-offs between objectives might eventually be necessary
    - Usability vs. authenticity
    - Structure vs. independency
    - Access vs. costs
    - …
- How can we ensure criteria are free of ambiguity?
The Objective Tree

• Tree structure describing requirements and goals
  • A weighted hierarchy of objectives leading into measurable criteria
  • A utility function for each criterion specifies the organisation’s assessment for the range of possible values

• Created top-down or bottom-up
  – Start from high-level goals and break down to specific criteria
  – Collect criteria and organize in tree structure
Influence Factors

Technology

Standards

User requirements

Characteristics of digital objects

format properties

other objectives

Goals and objectives

object properties

preservation process objectives

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
- Producers
- Others
- Administration
- Managers
- Lawyers
- Technical experts
- Consumers
An Objective Tree
Analog...
... or born-digital
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
- Yes/No (Y/N)
- Ordinal: define the possible values
- *(Subjective 0-to-5)*
Models for decision factors

- Format Properties
  - Library of Congress format evaluation
    http://www.digitalpreservation.gov/formats/
  - PRONOM format evaluation

- Software Quality
  - ISO 25010 SQUARE: Standardised software quality model

- Object properties
  - Formats and their properties
  - Representation Instances
  - Authenticity: Significant Properties
User perspective

- Goal of digital preservation is to serve (future) users in providing usable and authentic information
  - What is the preservation intent?
    
    http://www.dlib.org/dlib/january13/webb/01webb.html

- What are needs/requirements of users?
  - easy access
  - knowledge about origin of documents/ to be able to interpret them
  - to use them to 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)
  - …
Significant properties of ‘digital objects’

- What needs to be preserved?
  - Authenticity
  - Reliability
  - Integrity
  - Usability
  - Accuracy
  - Content
  - Context
  - Structure
  - Appearance
  - Behaviour
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…

Website

Technical characteristics
- Hardware
- Software
- Staff
- Automation
- Performance
- Documentation
- Usability
- Validation

Infrastructure characteristics

Process characteristics

Record characteristics

Appearance
- Fonts
- Page layout
- Highlighting

Content
- Documents
- Images
- Internal links present and working
- Tables
- Pagination
- Document structure

Structure

Behaviour

Context
File format characteristics

Support

Ubiquity

Quality

Disclosure

Openness

Availability

Documentation

Speed of change

Stability

Backwards compatibility

Ease of identification

Ease of validation

Lossiness

IPR

Complexity

Error tolerance

Comparative size

Larger/Same/Smallern

None/Detectable/Recoverable

Lossy/Lossless

Automatic/Manual/No

Automatic/Manual/No

None/Previous version only/Some previous versions/All previous versions

<1 year/1-2 years/3-5 years/>5 years

Public/Limited/None

Full/Partial/None

Standard/Open/Proprietary

Ubiquitous/Widespread/Specialised/Obsolete
• 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

Object characteristics

35% menus and navigation path

Y/N 35% complete 25% structure

Y/A/N 30% overall page layout

animated 25% mouse
pointer

speed 20% transitions
effects

speed 15% animations
colour

gradient 5% background

25% menu speed
type

colour
style

size 10% fonts

10% navigation

15% behaviour 80% reaction to activity

10% video/sound control

20% documentation material

10% context 80% metadata reference valid

40% content

Loops
Effects
content identical
quality
sound
picture
synchronisation

22% sound

22% video

22% image

22% text

12% user manual

FACULTY OF INFORMATICS
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>
Decision criteria: What to measure?

- Each criterion concerns either the action or its outcome

**Outcome**
- **Object** (authenticity, editability, …)
- **Format** (licensing, standardisation, complexity…)
- **Effect** (Costs…)

**Action**
- **Runtime** properties (performance, stability, logging…)
- **Static** (price, license…)
- **Judgement** (configuration interface usability…)

![Decision criteria diagram]
Results of Phase 1

- Defined and documented the context of a preservation problem
  - Which types of objects
  - Which environment
  - Purpose and target consumers
  - Obligations and constraints

- Defined and documented representative samples for performing experiments

- Defined and documented goals and objectives
  - From goals and requirements to measurable criteria
Questions?

... And a little bit of **READING for next week:**
- “Oh, you wanted us to preserve that?!“
- http://www.dlib.org/dlib/january13/webb/01webb.html

- Next: Presentation of the exercise task 2
DPUE Task 2: Project topics

- A number of topics published in TUWEL
  - Most are for groups of two people
  - Contact point for each
  - Make sure to get in touch and discuss!

- A few *musts*:
  - Sign up until the 22nd
  - Concept must be submitted until the 29th
  - Final results must be submitted until the 6th of June and
  - Final results must be presented
    on the 10th or *(read: and)* 17th of June.

- The presentation and ensuing discussions are integral part
  of this exercise and the lecture
  - You have to be present
DPUE Task 2: Project topics

1. Improve quality and coverage of characterisation (fits)
2. Reduce conflicts with post-processing rules in c3po
3. Progressive content profiling
4. Representative sample sets for experimentation
5. Interactive charting of content profiles
6. Simulation – estimating the next migration
7. Solution for preserving classical concert recordings
8. Optimizing decision making efficiency in planning
9. Data visualisation and analysis: Decision criteria
10. Quality assurance workflows
11. Resilient Web Services framework
12. Resilient Web Services register
13. Phaidra: Preservability assessment
14. Validation+verification of a migrated business process
15. Evaluation of processes (capturing measures)
16. Generation of a process context model
17. Recording+playback of user input for testing
In planning, a large number of decision criteria are evaluated.
  – These have different weighting and utility functions. Only after filling out all of them does the decision maker see the composite effect.
  – This is a quite expensive procedure, and the question arises if it can be done more efficiently. Can some of the criteria be left out from evaluation? Which? What is the effect of not providing a measure for them?

Goal: Develop a heuristic that computes potential savings in a preservation plan.
  – Calculate how many evaluation values could be left out
  – Would this change your confidence into the trustworthiness of the planning process?

Contact: Michael Kraxner (kraxner@ifs.tuwien.ac.at)
Thank you for your attention!

Don’t forget the easy **READING** for next week!

- “Oh, you wanted us to preserve that?!”
- [http://www.dlib.org/dlib/january13/webb/01webb.html](http://www.dlib.org/dlib/january13/webb/01webb.html)
- Next week:
  - How to plan: evaluation in Plato (hands-on)
  - Case studies: Examples of planning

[www.ifs.tuwien.ac.at/~becker](http://www.ifs.tuwien.ac.at/~becker)
[www.ifs.tuwien.ac.at/dp/plato](http://www.ifs.tuwien.ac.at/dp/plato)