Evaluating Preservation Strategies
for Electronic Theses and Dissertations

Stephan Strodl[1], Christoph Becker[1], Robert Neumayer[1], Andreas Rauber[1], Eleonora Nicchiarelli Bettelli[2], Max Kaiser[2], Hans Hofman[3], Heike Neuroth[4], Stefan Strathmann[4], Franca Debole[5] and Giuseppe Amato[5]

[1] Vienna University of Technology, Vienna, Austria
{strodl,becker,neumayer,rauber}@ifs.tuwien.ac.at
[2] Austrian National Library, Vienna, Austria
{eleonora.nicchiarelli,max.kaiser}@onb.ac.at
[3] Nationaal Archief, The Hague, The Netherlands
hans.hofman@nationaalarchief.nl
[4] State and University Library Goettingen, Goettingen, Germany
{neuroth,strathmann}@mail.sub.uni-goettingen.de
[5] Italian National Research Council (CNR), Pisa, Italy
{franca.debole,giuseppe.amato}@isti.cnr.it

Abstract

Digital preservation has turned into a pressing challenge for institutions having the obligation to preserve digital objects over years. A range of tools exist today to support the variety of preservation strategies such as migration or emulation. Yet, different preservation requirements across institutions and settings make the decision on which solution to implement very difficult. The Austrian National Library will have to preserve electronic theses and dissertations provided as PDF files and is thus investigating potential preservation solutions. The DELOS Digital Preservation Testbed is used to evaluate various alternatives with respect to specific requirements. It provides an approach to make informed and accountable decisions on which solution to implement in order to preserve digital objects for a given purpose. We analyse the performance of various preservation strategies with respect to the specified requirements for the preservation of master theses and present the results.

Categories and Subject Descriptors
H.3 Information Storage and Retrieval: H.3.7 Digital Libraries;

General Terms
Digital Library, Digital Preservation, Long Term Access

Keywords
Preservation Planning, Migration, Emulation, Case Study

1   Introduction

An increasing number of organisations throughout the world face national as well as institutional obligations to collect and preserve digital objects over years. To fulfil these obligations the institu-

tions are facing the challenge to decide which digital preservation strategies to follow. This selection of a preservation strategy and tools is the most difficult part in digital preservation endeavours. The decision depends on the institutional needs and goals for given settings. Technical as well as process and financial aspects of a preservation strategy form the basis for the decision on which preservation strategy to adopt.

A number of strategies have been devised over the last years. An overview is provided by the companion document to the UNESCO charter for the preservation of the digital heritage (UNESCO, 2003). All of the proposed strategies have their advantages and disadvantages, and may be suitable in different settings. The most common strategy at the moment is migration, where the object is converted into a more current or more easily preservable file format such as the recently adopted PDF/A standard (ISO, 2004), which implements a subset of PDF optimised for long-term preservation. A report about different kinds of risks for a migration project is done by the Council of Library and Information Resources (CLIR) (Lawrence et al., 2000). Another important strategy is emulation, which aims to provide programmes that mimic a certain environment, e.g. a certain processor or the features of a certain operating system. Jeff Rothenberg (Rothenberg, 1999) envisions a framework of an ideal preservation surrounding. PANIC (Hunter and Choudhury, 2005) addresses the challenges of integrating and leveraging existing tools and services and assisting organisations to dynamically discover the optimum preservation strategy.

The Austrian National Library (ONB) will have the future obligation to collect and preserve electronic theses and dissertations from Austrian universities. To fulfil this obligation, the ONB needed a first evaluation of possible preservation strategies for these documents according to their specific requirements.

The DELOS DP Testbed allows the assessment of all kinds of preservation actions against individual requirements and the selection of the most suitable solution. It enforces the explicit definition of preservation requirements and supports the appropriate documentation and evaluation by assisting in the process of running preservation experiments. Thus it was used for assessing potential strategies.

The approach presented in this paper basically focuses on the elicitation and documentation of the requirements (objectives). File format repositories such as PRONOM (Pettitt, 2003) may be used to identify specific technical characteristics of the digital objects at hand.

In this paper we describe the workflow for evaluating and selecting DP solutions following the principles of the DELOS DP Testbed. We present the results of the case study involving the Austrian National Library, and demonstrate the benefits of the proposed approach.

The remainder of this paper is organised as follows: Following an overview of the principles of the DELOS DP Testbed in Section 2, a description of the workflow is presented in Section 3. We report on the case study on the preservation of electronic theses in PDF format in Section 4. An overview of other case studies is given in Section 5. The closing Section 6 provides conclusions, lessons learned as well as an outlook on future work.

2 The DELOS DP Testbed

The DELOS DP Testbed of the DELOS Digital Preservation Cluster combines the Utility Analysis approach (Rauch and Rauber, 2004) with the testbed designed by the Dutch National Archive. Figure 1 provides an overview of the workflow of the DELOS DP Testbed, which was described in (Rauch and Rauber, 2004) and recently revised and described in detail in (Strodl et al., 2006). The 3-phase process, consisting of 14 steps, starts with defining the scenario, setting the boundaries, defining and describing the requirements to be fulfilled by the possible alternatives of preservation actions. The second part of the process identifies and evaluates potential alternatives. The alternatives' characteristics and technical details are specified; then the resources for the experiments are selected, the required tools set up, and a set of experiments is performed. Based on the requirements defined in the beginning, every experiment is evaluated. In the third part of the workflow the results of the experiments are aggregated to make them comparable, the importance factors are set, and the alternatives are ranked. The stability of the final ranking is analysed with respect to minor changes in the weighting and performance of the individual objectives using Sensitivity Analysis. The results are finally evaluated by taking non-measurable influences on the decision into account. After this analysis a clear and well argumented accountable, recommendation for one of the alternatives can be made.

To simplify the process, to guide users and to automate the structured documentation, a software tool is introduced[6]. It implements the workflow of the DELOS DP Testbed, supporting the documentation of the various steps performed. Results may be stored centrally on a server or exported to an XML file.

3 Testbed Workflow

The detailed workflow consists of fourteen steps as shown in Figure 1, which are described below.

1. **Define Basis**
   The basis of the DELOS Digital Preservation Testbed is a semi-structured description including the required types of records to be considered, a description of the environment in which the testbed process takes place, and information on the amount of files or records.
2. **Choose Records**
   This step selects sample records representing the variety of document characteristics of the considered collection. These samples are later used for evaluating the preservation alternatives.
3. **Define Requirements**
   The goal of this decisive step is to clearly define the requirements and goals for a preservation solution in a given application domain. In the so-called objective tree, high-level goals and detailed requirements are collected and organised in a tree structure.
   While the resulting trees usually differ through changing preservation settings, some general principles can be observed. At the top level, the objectives can usually be organised into four main categories:
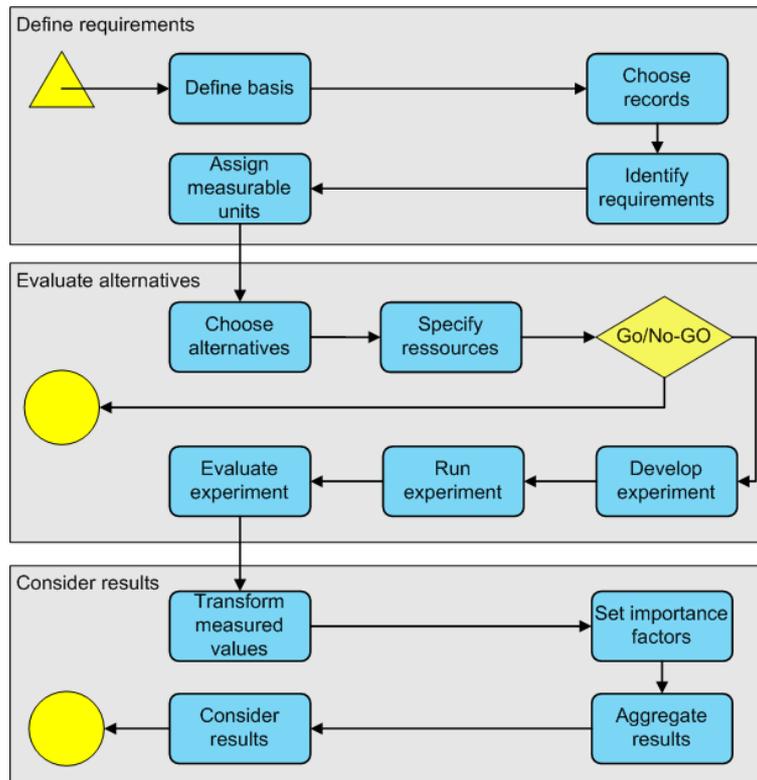
---

[6] http://ifs.tuwien.ac.at/dp

Fig. 1. Overview of DELOS Digital Preservation Testbed's workflow

- *File characteristics* describe the visual and contextual experience a user has by dealing with a digital record. Subdivisions may be "Appearance", "Content", "Structure" and "Behaviour", with lowest level objectives being e.g. colour depth, image resolution, forms of interactivity, macro support, or embedded metadata.
- *Record characteristics* describe the technical foundations of a digital record, the context, the storage medium, interrelationships and metadata.
- *Process characteristics* describe the preservation process. These include usability, complexity or scalability.
- *Costs* have a significant influence on the choice of a preservation solution. Usually, they may be divided in technical and personnel costs.

The objective tree is usually created in a workshop setting with experts from different domains contributing to the requirements gathering process. The tree documents the individual preservation requirements of an institution for a given partially homogeneous collection of objects. Examples include scientific papers and dissertations in PDF format, historic audio recordings, or video holdings from ethnographic studies. Typical trees may contain between 50 to several hundred objectives, usually organised in 4-6 hierarchy levels.

4. **Assign Measurable Units**
   Measurable effects are assigned to the objectives that have been defined in the previous step.

Wherever possible, these effects should be objectively measurable (e.g. € per year, frames per second). In some cases, (semi-) subjective scales will need to be employed (e.g. degrees of openness and stability, support of a standard, degree of file format adoption, etc.).

5. **Choose Alternatives**
   Different preservation solutions, such as different migration tools or emulators, are described. An extensive description of the preservation process ensures a clear understanding of each alternative.

6. **Specify Resources**
   For each alternative defined in the previous step, a project and work description plan is developed, where the amount of work, time and money required for testing the alternative are estimated.

7. **Go/No-Go**
   This step considers the definition of resources and requirements to determine if the proposed alternatives are feasible. The result is a decision for continuing the evaluation process or a justification of the abandonment of certain alternatives.

8. **Develop Experiment**
   In order to run repeatable tests, a documented setting is necessary. This stage produces a specific development plan for each experiment, which includes the workflow of the experiment, software and hardware system of the experiment environment, and the mechanism to capture the results.

9. **Run Experiment**
   An experiment will test one or more aspects of applying a specific preservation alternative to the previously defined sample records.

10. **Evaluate Experiments**
    The results of the experiments are evaluated to determine the degree to which the requirements defined in the objective tree were met.

11. **Transform Measured Values**
    The measurements taken in the experiments might follow different scales. In order to make these comparable, they are transformed to a uniform scale using transformation tables. The resulting scale might e. g. range from 0 to 5. A value of 0 would in this case denote an unacceptable result and thus serve as a drop-out criterion for the whole preservation alternative.

12. **Set Importance Factors**
    Not all of the objectives of the tree are equally important. This step assigns importance factors to each objective depending on specific preferences and requirements of the project.

13. **Aggregate Results**
    With the input of the importance factors and the transformed numbers, a single final value of each alternative is calculated.

14. **Perform Sensitivity Analysis and Consider Results**
    Finally the alternatives are ranked. The software implementation supports varying weights form different users. These are further used for the Sensitivity Analysis of the evaluation, which analyses, e. g., the stability of the ranking with respect to minor changes in weighting of the individual objectives. Additionally, side effects may be considered that are not included in the numerical evaluation, like the relationship with a supplier or expertise in a certain alternative.

4 Preserving Austrian Theses and Dissertations

The Austrian National Library will have the future obligation to collect and preserve master theses from Austrian Universities. The theses will be provided to the library in a PDF format. The Austrian National Library provides guidelines for creating preservable PDFs (Horvath, 2005), but at the moment the ONB is not able to legally enforce these guidelines. This case study gives a starting point to identify the requirements and goals for the digital preservation of master theses. It furthermore allows a first evaluation of the various preservation actions being considered.

This application domain is interesting and highly relevant for digital preservation practice for a number of reasons:

1. PDF is a wide-spread file format and very common in libraries and archives.
2. Although PDF is a single file format, there exist different versions of the standard.
3. Different embedded objects are captured in this case study, such as video and audio content.

In a brainstorming workshop the requirements for this specific application area were collected. The resulting objective tree shows a strong focus on the structure, content and appearance of the objects; especially layout and structure of the documents need to be preserved.

Characteristics concerning object structure include among others

- Document structure (chapters, sections),
- Reference tables (table of content, list of figures)
- Line and page breaks,
- Headers and footers,
- Footnotes,
- Equations (size, position, structure, caption),
- Figures (size, position, structure, caption), and
- Tables (size, position, structure, caption).

The next step was to assign measurable effects for each leaf of the tree. Most of them are simple yes/no decisions, for example whether the fontsize of text changed or not, or whether table structures have been kept intact.

The weighting of the tree reflects the primary focus on content; at the top level the object characteristics as well as process characteristics and costs have a strong influence on the choice of a preservation strategy.

Several migration solutions were evaluated using the DELOS DP Testbed:

1. Conversion to plain-text format using Adobe Acrobat 7 Professional.
2. Conversion to Rich Text Format (RTF) using SoftInterface ConvertDoc 3.82.
3. Conversion to RTF using Adobe Acrobat 7 Professional.
4. Conversion to Multipage TIFF using Universal Document Converter 4.1.

Nr	Alternative	Total score
1	PDF/A (Adobe Acrobat 7 prof.)	4.52
2	TIFF (Document Converter 4.1)	4.26
3	EPS (Adobe Acrobat 7 prof.)	4.22
4	JPEG 2000 (Adobe Acrobat 7 prof.)	4.17
5	RTF (Adobe Acrobat 7 prof.)	3.43
6	RTF (ConvertDoc 4.1)	3.38
7	TXT (Adobe Acrobat 7 prof.)	3.28

Table 1. Overall scores of the alternatives

5. Conversion to PDF/A using Adobe Acrobat 7 Professional.
   (The generated PDF/A is not completely consistent with PDF/A-ISO-Standard. (ISO, 2004))
6. Conversion to lossless JPEG2000 using Adobe Acrobat 7 Professional.
7. Conversion to Encapsulated PostScript (EPS) using Adobe Acrobat 7 Professional.

All experiments were executed on Windows XP professional on a sample set of five master theses from the Vienna University of Technology. The results as provided in Table 1 show that the migration to PDF/A using Adobe Acrobat 7 Professional ranks on top, followed by migration to TIFF, EPS and JPEG2000; far behind are RTF and plain text. The alternative PDF/A basically preserves all core document characteristics in a wide-spread file format, while showing good migration process performance.

The alternatives TIFF, EPS and JPEG show very good appearance, but have weaknesses regarding criteria such as 'content machine readable'. Furthermore, the migration to JPEG and EPS produces one output file for each page, the object coherence is not as well preserved as in a PDF/A document.

Both RTF solutions exhibit major weaknesses in appearance and structure of the documents, specifically with respect to tables and equations as well as character encoding and line breaks. Object characteristics show a clear advantage for ConvertDoc, which was able to preserve the layout of headers and footers as opposed to Adobe Acrobat. Still, costs and the technical advantages of the Acrobat tool, such as macro support and customization, compensate for this difference and lead to an equal score.

The loss of essential characteristics means that the plain text format fails to fulfil a number of minimum requirements regarding the preservation of important artifacts like tables and figures as well as appearance characteristics like font types and sizes.

Figure 2 shows an extract of the results of the software tool. The software supports the comparison of alternatives to highlight strengths and weaknesses of each alternative.

Multimedia content proved to be a difficult task: None of the tested alternatives was able to preserve embedded audio and video content. This issue could be solved in two ways: (1) Use a tool for automated extraction of multimedia content from PDF. (2) Solve the problem on an organisational level by issuing a submission policy which states that multimedia objects have to be provided separately. In both cases, a separate preservation strategy for the multimedia content has to be devised.
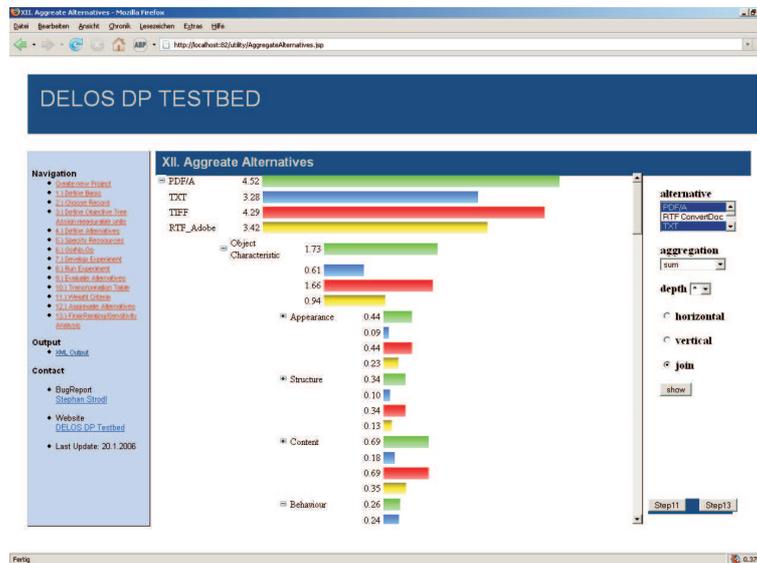
Fig. 2. Screenshot: comparison of results

Depending on whether preserving multimedia content is a primary goal to be fulfilled, our final recommendation resulting from the evaluation of the experiments is to (1) use migration to PDF/A with Adobe Acrobat 7 Professional or (2) combine the alternative PDF/A with a multimedia extraction tool or a submission policy.

5 Case Studies

Further case studies were conducted within DELOS during the last years include the following.

– **Video Files of the Austrian Phonogrammarchiv**
 The Austrian Phonogrammarchiv is re-considering its appraisal regulations for video files, specifically with respect to optimal source format standards to migrate from. So a case study took place to evaluate the performance of potential migration tools and source formats. The defined target format was MPEG2000 and DPS, by considering all occurring input formats (Std DVm Digi-Betam PAL-VHS, SVHS, U-Matic, Beta Cam, MPEG, NTSC-VHS, DPS, Hi8). In a one-day workshop an objective tree was created with around 200 objectives. These were strongly focused on detailed technical characteristics. The subsequent experiments and the evaluation of the preservation solutions took about 3 weeks. The results clearly revealed the few distinguishing characteristics of the alternative preservation strategies – signal representation, colour depth and stereo quality.

– **Document records of the Dutch National Archive**

The Dutch National Archive is responsible for storing all documents generated by the Dutch government, ministries and official bodies. The case study tried to define the objectives for the preservation of different kinds of documents, such as video and audio, focusing particularly on the record characteristics. The resulting objective tree contained around 450 objectives.

– **Migration of a database to XML**

This case study was done in cooperation with the Italian National Research Council (CNR). The starting point was a legacy database containing descriptive meta data of a small library, consisting of books, registered users, information about lending, order of books, content (field, review) and the budget for new books. The data of the database was to be converted in XML for archiving and further application using e.g. a native XML database. In this case study we tried to reduce the number of objectives, focusing on the critical characteristics. The resulting objective tree contained approximately 70 nodes with a maximum depth of 6 layers.

– **Preserving annual electronic journal of differential equations**

Within the project of supra-regional literature supply in Germany the State and University Library Goetting holds the collection of "Electronic Journal of Differential Equations". The SUB is committed to preserver the collection of the journals and providing access to them. In a first workshop the requirements and goals for the collection were specified. The specific challenges of this collection are the hierarchical structure of the considered object and the different formats of the sub-objects.

6   Conclusions

The DELOS DP Testbed provides a means to make well-documented, accountable decisions on which preservation solution to implement. It enforces the explicit definition of preservation requirements in the form of specific objectives. It allows to evaluate various preservation solutions in a consistent manner, enabling informed and well-documented decisions. Thus, it helps to establish and maintain a trusted preservation environment.

The case study of the Austrian National Library evaluates various migration strategies for PDF. The migration to PDF/A by Adobe Acrobat 7 Professional reaches the highest score and provides a feasible solution for the long term storage of theses and dissertations. Migration to TIFF, EPS and JPEG perform very good at appearance objectives, but have some substantial technical weaknesses. The preservation alternatives RTF and plain text are not able to migrate essential parts of the object and should not be considered further. None of the evaluated alternatives is able to handle multimedia content, this issue has to be solved on another appropriate level - either by extracting the multimedia content or by issuing a submission policy. Further work will evaluate different tools for converting PDF to PDF/A with a focus on process objectives such as duration, capacity, and automation support.

While many of the processing steps of the DELOS DP Testbed are automated, a significant amount of work is still involved in acquiring the measurements of the experiment outcomes. Ongoing work on preservation plan decision support within the European Union PLANETS project (http://www.planets-project.eu) is based on the DELOS Digital Preservation Testbed. It will integrate tools for preservation action and object characterisation to further reduce the workload.

Acknowledgements

Bibliography

[Horvath, 2005]Horvath, M.: 2005, *Empfehlungen zum Erzeugen archivierbarer Dateien im Format PDF*, Technical report, Austrian National Library, `http://www.onb.ac.at/about/lza/pdf/ONB_PDF-Empfehlungen_1-4.pdf`(in German)

[Hunter and Choudhury, 2005]Hunter, J. and Choudhury, S.: 2005, in *Proceedings of the 5th International Web Archiving Workshop (IWAW05)*, Vienna, Austria

[ISO, 2004]ISO: 2004, *Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A) (ISO/CD 19005-1)*, International Organization for Standardization

[Lawrence et al., 2000]Lawrence, G., Kehoe, W., Rieger, O., Walters, W., and Kenney, A.: 2000, *Risk Management of Digital Information: A File Format Investigation*, Technical report, Council on Library and Information Resources

[Pettitt, 2003]Pettitt, J.: 2003, *PRONOM - Field Descriptions*, The National Archives, Digital Preservation Department, `http://www.records.pro.gov.uk/pronom`

[Rauch and Rauber, 2004]Rauch, C. and Rauber, A.: 2004, in *Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL 2004)*, pp 203–212, Springer

[Rothenberg, 1999]Rothenberg, J.: 1999, *Avoiding Technological Quicksand: Finding a viable technical foundation for digital preservation*, Technical report, Council on Library and Information Resources

[Strodl et al., 2006]Strodl, S., Rauber, A., Rauch, C., Hofman, H., Debole, F., and Amoato, G.: 2006, in *Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL 2006)*, pp 323–332, Springer

[UNESCO, 2003]UNESCO: 2003, *Guidelines for the preservation of digital heritage*, UNESCO, Information Society Division, `http://www.unesco.org/webworld/mdm`