Long-Term Preservation of Electronic Theses and Dissertations: A Case Study in Preservation Planning ♣

© Christoph Becker[1], Stephan Strodl[1], Robert Neumayer[1], Andreas Rauber[1],
Eleonora Nicchiarelli Bettelli[2], Max Kaiser[2]
[1] Vienna University of Technology, Vienna, Austria
{becker,strodl,neumayer,rauber}@ifs.tuwien.ac.at
[2] Austrian National Library, Vienna, Austria
{eleonora.nicchiarelli,max.kaiser}@onb.ac.at

Abstract

An increasing number of institutions throughout the world face legal obligations to collect and 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 in PDF. It is thus investigating potential preservation solutions. The preservation planning approach taken in the PLANETS project 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's theses and dissertations and present the results.

1. Introduction

Digital objects have become the dominant way that we create, shape, and exchange information. They increasingly contain essential parts of our cultural, intellectual and scientific heritage; they form a central part of our economy, and increasingly shape our private lives. The ever-growing heterogeneity and complexity of digital file formats together with rapid technological changes turn the preservation of digital information into a pressing challenge. The challenge is to keep electronic data accessible, viewable, and usable for the future, to ensure the survival of our digital artefacts when the original software or hardware to interpret them correctly becomes unavailable [13].

Proceedings of the 9th Russian Conference on Digital Libraries RCDL'2007, Pereslavl, Russia, 2007

Digital preservation deals with the long-term storage and access to digital objects. The Digital Preservation Coalition defines it as *the series of managed activities necessary to ensure continued access to digital materials* and adds that it *refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change [5]*. It is confined from digitisation, which is a challenging field in itself.

Over the last years, a lot of effort was spent on defining, improving, and evaluating preservation strategies. All of the proposed strategies have their advantages and disadvantages, and may be suitable in different settings. A good overview of preservation of digital heritage and preservation strategies is provided by the companion document to the UNESCO charter for the preservation of the digital heritage [13].
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 [4], 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) [6].
Another important strategy is **emulation**, which aims at providing programs that mimic a certain environment, e.g. a certain processor or the features of a certain operating system. Jeff Rothenberg [10] envisions a framework of an ideal preservation surrounding.
The Universal Virtual Computer (UVC) concept [1] uses elements of both migration and emulation, allowing digital objects to be reconstructed in their original appearance. The UVC is independent of any existing hardware or software; it simulates a basic architecture including memory, register and rules. In the future only a single emulation layer between the UVC and the computer is necessary to reconstruct a digital object in its original appearance.

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 institutions 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.

Several projects in this domain have been initiated under the 6th framework program of the European Union. One of the main objectives of the DELOS Digital Preservation Cluster[1], which is part of the EU-funded DELOS Network of Excellence on Digital Libraries, is the establishment of testbeds and validation metrics. The DELOS Digital Preservation Testbed [11] allows the selection of the most suitable preservation strategy for individual requirements by combining a structured workflow for requirements specification and evaluation by means of a standardised testbed laboratory infrastructure.

PLANETS (Permanent Long-term Access through Networked Services)[2] develops systems and tools which support the accessibility and use of digital cultural and scientific resources. More specifically, the project is developing methods and tools based on a distributed service infrastructure on which services for preservation action, preservation characterisation and preservation planning can be coordinated and combined with each other. The PLANETS Testbed will use this framework to provide a stable foundation for evaluating different preservation actions in a well-defined setting. A central part of PLANETS is the methodology for preservation planning which is based on the results of the DELOS project.

The shift to electronic publishing in the last decade has also influenced the way theses and dissertations are created and collected. Electronic theses and dissertations (ETD) are now the standard. The Networked Digital Library of Theses and Dissertations (NDLTD)[3] has organised a series of workshops on this topic.
McMillan et.al. [7] give an overview of published ETD initiatives and describe a digital preservation network for ETDs based on LOCKSS[4]. Dobratz [1] presents an XML-based publishing system for handling ETDs that focuses on open access and long-term preservation.

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 Preservation Planning approach in the PLANETS project 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 [8] 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 digital preservation solutions following the principles of the PLANETS preservation planning methodology. 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 PLANETS approach to Preservation Planning 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. The closing Section 5 provides conclusions, lessons learned as well as an outlook on future work.

2 PLANETS Preservation Planning

The preservation planning methodology of the PLANETS project is based on the DELOS DP Testbed developed as part of the DELOS Digital Preservation Cluster, which combines Utility Analysis with the Testbed designed by the Dutch National Archive. The DELOS Testbed was described in [9] and recently revised and described in detail in [11]. Figure 1 provides an overview of the PLANETS Preservation Planning workflow. See [12] for a detailed description.
The 3-phase process, consisting of 11 steps, starts with defining the scenario and boundaries and specifying 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, well argued and accountable recommendation for one of the alternatives can be made.
As part of the PLANETS project, we are developing a decision support tool that supports and automates the

[1] http://www.dpc.delos.info
[2] http://www.planets-project.eu
[3] http://www.ndltd.org
[4] http://www.lockss.org

preservation planning process as described below. The planning tool will be integrated into a distributed service architecture that allows a user to dynamically discover, execute and evaluate preservation services such as migration tools.

3 Preservation Planning Workflow

The workflow consists of eleven steps, which are described in the following and shown in Figure 1.

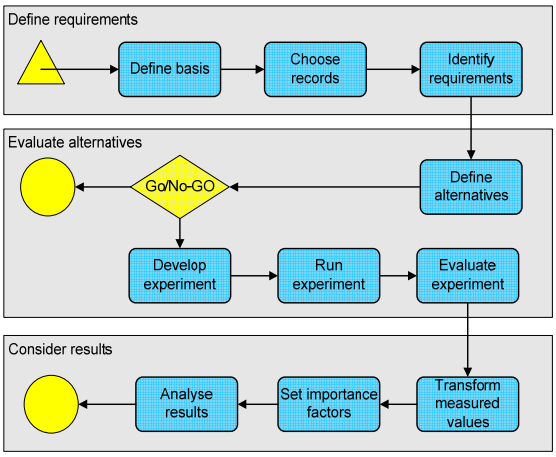

Figure 1 Overview of the PLANETS Preservation Planning workflow

Step 1 - Define basis

The basis of a preservation plan is a semi-structured description including the required types of records to be considered, a description of the environment in which the planning process takes place, and information on the amount of files or records.

Step 2 - Choose records

In order to be able to evaluate alternative strategies, sample records are chosen to run the experiments. These sample records have to be representative for the collection that is being considered. They should exhibit all the specific peculiarities and essential properties of objects and thus cover the variety of the elements in the collection.

Step 3 – Identify 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 in preservation settings, some general principles can be observed. At the top level, the objectives can usually be organised into four main categories:

- **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”.
- **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.

Typical trees may contain from about 50 up to several hundred objectives, usually organised in 4-6 levels of hierarchy.
Measurement units are assigned to each leaf in the tree. Wherever possible, these measurements should be objectively measurable (e.g. seconds per objects, dots-per-inch resolution, etc.). In some cases, (semi-) subjective scales will need to be engaged, e.g. degrees of openness and stability, support of a standard, etc.

Step 4 – Define alternatives

Different alternative strategies need to be defined which should be evaluated in the planning tool. The preservation planning software will support this step by listing available preservation action services obtained from a service registry.
In order to assess the resources that are needed to run the evaluation, the planner estimates the amount of work, time and money for each potential alternative.

Step 5 - Go/No-Go decision

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 for the abandonment of certain alternatives.

Step 6 - Develop experiment

A documented setting is necessary in order to run repeatable tests. 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. The experiments can be either performed by using online services or by using an environment external to the planning tool.
In the case of online services the planner sets parameters and may additionally select services for capturing the results. In external environments, all software packages and programs needed for the experiment will be developed and/or installed and tested, including copies of all the objects needed for the experiment and mechanisms for capturing the results.

Step 7 - Run experiment

An experiment tests one or more aspects of applying a specific preservation alternative to the previously

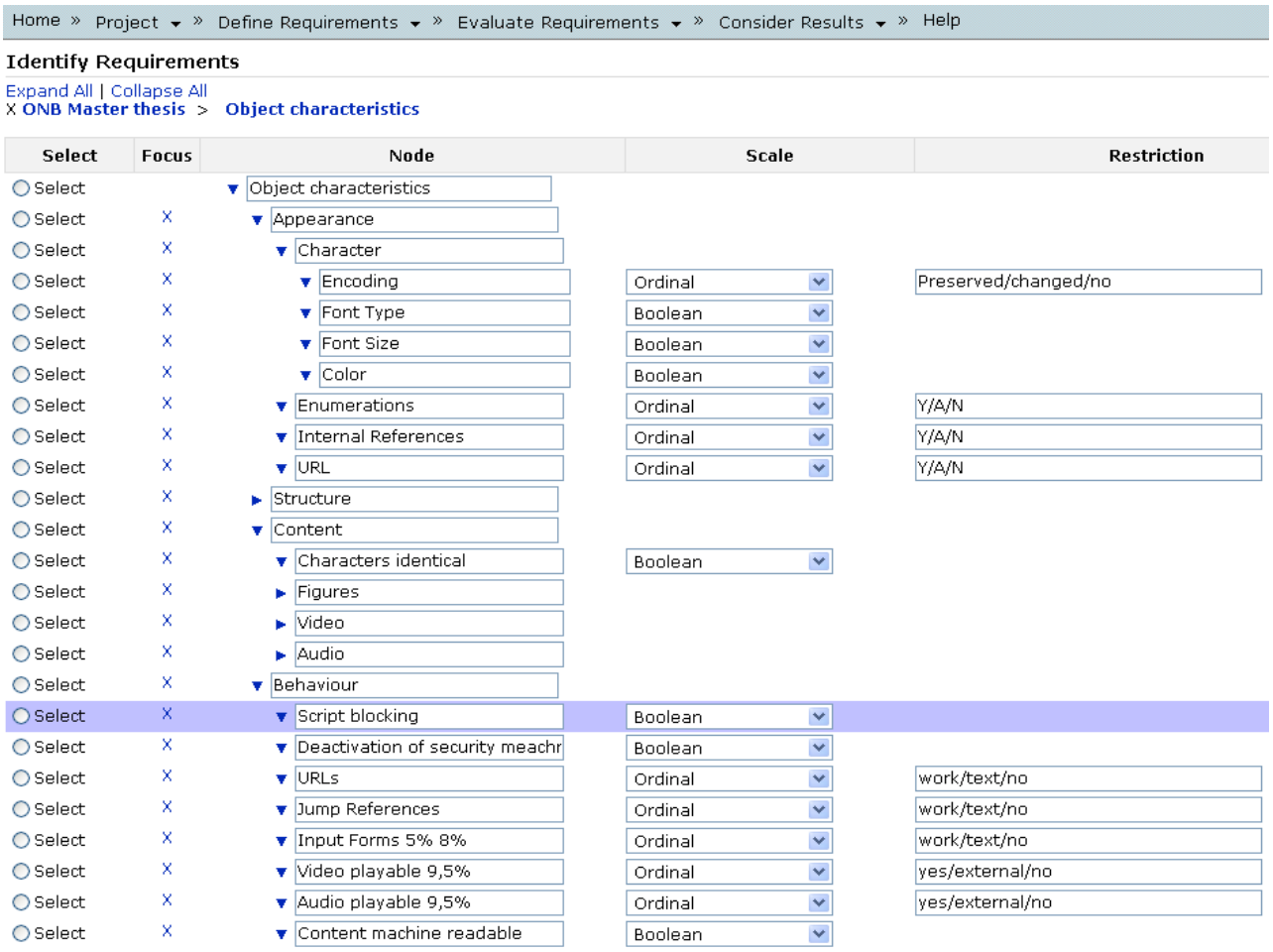

Figure 2 Screenshot of the Preservation Planning software

defined sample records. Running an experiment will produce results that will be evaluated in the next step.

Step 8 - Evaluate experiment

The results of the experiments are evaluated to determine the degree to which the requirements defined in the objective tree were met. The evaluation can be supported by online characterisation services.

Step 9 - Transform values

The measurements taken in the experiments might all have different scales. In order to make these comparable, they are transformed to a uniform scale using transformation functions.

Step 10 - Set importance factors

Not all of the objectives are equally important. This step assigns relative importance factors in a top down manner to each objective to explicitly describe which objectives play a major or minor role for the final decision. These weights depend largely on individual requirements.

Step 11 – Analyse Results

In this step the performance measures for the individual objectives are aggregated to one single comparable number for each alternative. The measured performance values are transformed by the transformation functions and multiplied with the weighting factor. These numbers are aggregated to a single comparable number per alternative. We thus obtain performance values for each part of the objective tree and for each alternative, including an overall performance value at the root level. Several aggregation methods are available; the most relevant are the following:

1. *Weighted Sum* – This standard method results in a rational scale.
2. *Weighted Multiplication* – Also on a rational scale, this method is distinguished from the weighted sum in that knock-out criteria that are mapped to a performance value of 0 are propagated throughout the tree, thus indicating that an alternative fails to fulfil essential requirements.

4 Preserving Electronic 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 PDF format. The Austrian National Library provides guidelines for creating preservable PDFs [3], 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 widely adopted 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 elicitation and definition of requirements during a brainstorming session, as well as the subsequent structuring to form an objective tree, were initially performed in a traditional manner, using staples of post-it notes on a white-board. During the course of the PLANETS project, this situation has been greatly improved by using mind-mapping software to construct the tree and importing the resulting XML definition into the planning software. Figure 2 depicts the editor for the resulting tree in the decision support tool.

The next step is to assign measurable effects for each leaf of the tree. Some are simple yes/no decisions, for example whether the font size of text changed or whether table structures have been kept intact. Others, such as performance of a migration tool, can be measured on a numeric scale.

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 PLANETS preservation planning methodology:

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.
5. Conversion to PDF/A using Adobe Acrobat 7 Professional. (Note that the generated PDF/A is not completely consistent with the PDF/A-ISO-Standard [4].)
6. Conversion to lossless JPEG2000 using Adobe Acrobat 7 Professional.
7. Conversion to Encapsulated PostScript (EPS) using Adobe Acrobat 7 Professional.
8. We also evaluated the alternative of not migrating at all, which leaves the documents in their original formats. As there are multiple versions of PDF, this of course incurs additional risks.

All experiments were executed on Windows XP professional on a sample set of five master's 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 widely adopted file format, while showing good migration process performance.

Note that while the option of leaving the documents in their original PDF format(s) seems to show good performance when looking at the overall weighted sum aggregation, weighted multiplication reveals that some essential requirements are not met. These are the deactivation of scripting and security mechanisms, which are regarded a knock-out criterion that must be fulfilled.

The alternatives TIFF, EPS and JPEG show very good appearance, but have weaknesses regarding criteria such as 'content machine readable'. Furthermore, as 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 customisation, compensate for this difference and lead to an equal score.

Alternative	Total Score Weighted Sum	Total Score Weighted Multiplication
PDF/A (Adobe Acrobat 7 prof.)	4.52	4.31
PDF (unchanged)	4.53	0.00
TIFF (Document Converter 4.1)	4.26	3.93
EPS (Adobe Acrobat 7 prof.)	4.22	3.99
JPEG 2000 (Adobe Acrobat 7 prof.)	4.17	3.77
RTF (Adobe Acrobat 7 prof.)	3.43	0.00
RTF (ConvertDoc 4.1)	3.38	0.00
TXT (Adobe Acrobat 7 prof.)	3.28	0.00

Table 1: Overall scores of the alternatives

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

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 organizational 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.

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 Conclusions

The proposed approach to preservation planning 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 evaluating 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. Leaving the objects untouched fails to deactivate active content and is thus not an option. Migration to TIFF, EPS and JPEG perform very well at appearance objectives, but have some substantial technical weaknesses because the contained text cannot be accessed easily anymore. The alternatives RTF and plain text are not able to preserve 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 there is considerable tool support guiding the requirements specification and decision process, a significant amount of work is still involved in acquiring the measurements of the experiment outcomes. Ongoing work within the PLANETS project will integrate tools for preservation action and object characterisation to further reduce the workload.

References

[1] Dobratz, S., 2005. *Thinking the long term: The XML-based publishing workflow for handling electronic theses and dissertations at Humboldt-University Berlin.* In Proceedings of Eighth Symposium on Electronic Theses and Dissertations (ETD 2005*)*. Sydney, Australia, 28-30 September 2005.

[2] Hoeven, J., Van Der Diessen, R., and Van En Meer, K., 2005. *Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects*. Journal of Information Science, Vol. 31 (3)

[3] 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).

[4] 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.

[5] Jones, M., and Beagrie, N., 2002. *Preservation Management of Digital Materials: A Handbook.* Digital Preservation Coalition, 2002. http://www.dpconline.org/graphics/handbook

[6] 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.

[7] McMillan, G., Jannik, C.M., and McDonald, R.H.: *A Practical, Working and Replicable Approach to ETD Preservation*. In Proceedings of Eighth Symposium on Electronic Theses and Dissertations (ETD 2005*)*. Sydney, Australia, 28-30 September 2005.

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

[9] Rauch, C. and Rauber, A. 2004. *Preserving digital media: Towards a preservation solution evaluation metric*. In Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL 2004). Shanghai (China), 13－17 December 2004. Springer, 203–212.

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

[11] Strodl, S., Rauber, A., Rauch, C., Hofman, H., Debole, F., and Amato, G. 2006. *The DELOS Testbed for Choosing a Digital Preservation Strategy*. In Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL 2006).Kyoto (Japan), 27-30. November 2006. Berlin-Heidelberg: Springer. 323–332.

[12] Strodl, S., Becker, C., Neumayer, R., Rauber, R. 2007. *How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure*. In: Proceedings of the ACM IEEE Joint Conference on Digital Libraries (JCDL'07), June 2007, Vancouver, British Columbia, Canada. 29-38

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

♣ Part of this work was supported by the European Union in the 6. Framework Program, IST, through the DELOS DPC Cluster (WP6) of the DELOS Network of Excellence on Digital Libraries, contract 507618, and the PLANETS project, contract 033789.