How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure

Stephan Strodl, Christoph Becker, Robert Neumayer, Andreas Rauber
Vienna University of Technology
Vienna, Austria
www.ifs.tuwien.ac.at/dp

ABSTRACT

An increasing number of institutions throughout the world face legal obligations or business needs to collect and preserve digital objects over several decades. Today, 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.

This paper presents the *PLANETS Preservation Planning approach*. It provides an approved way to make informed and accountable decisions on which solution to implement in order to optimally preserve digital objects for a given purpose. It is based on Utility Analysis to evaluate the performance of various solutions against well-defined requirements and goals. The viability of this approach is shown in several case studies for different settings. We present its application to two scenarios of web archives, two collections of electronic publications, and a collection of multimedia art. This work focuses on the different requirements and goals in the various preservation settings.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.7 Digital Libraries

General Terms

Measurement, Documentation, Performance, Experimentation

Keywords

Digital Libraries, Digital Preservation, Preservation Planning, OAIS Model, Utility Analysis, Evaluation

1. INTRODUCTION

Digital Preservation – the process of keeping electronic material accessible and usable for a certain period of time – has turned into one of the most pressing challenges within the digital library community. Not only because of the rapid changes and ongoing developments in file formats, long-term archiving of digital material is a highly complex and diverse matter. At the same time, technical advances occur in hardware development, and the information technology infrastructure as well as computer equipment are subject to changes. The ever-growing amount of material being available digitally not only drives the need for feasible access and delivery, but also for preserving digital objects in the medium and long run. A wide variety of institutions and individuals from countless fields have a strong interest in keeping their digital objects accessible and usable over the next decades. Private users who want to keep accessible their photo, audio, or video collections as well are a significant target audience. Insurance and aviation companies, the pharmaceutical and car industry, and other key players have a strong interest or obligations to preserve their data holdings, simulation models, or studies over time. Some institutions are required to keep copies of their documents due to legal constraints. For others, providing digital content is a central part of their business model and therefore the time aspect is a very important one as well. Libraries and museums are increasingly digitising their holdings or even holding born-digital content. These assets need to be preserved for future generations, forming an enormously wide range of material with a vast number of differing requirements. Moreover, electronic content has to be monitored to guarantee the accessibility and usability over time. The combined complexity of these facts is one of the main challenges of digital preservation as a research discipline.

The wide and constantly growing variety of file formats currently available ranges from simple formats like plain text, which consists of simple ASCII or unicode characters, to more complex or compound file types. PDF or Microsoft Word files, for instance, can contain additional formatting information or images and tables, all of which have special properties. Going further, multimedia presentations add to this complexity the characteristics of all kinds of embedded objects like image, video, or audio files.

Individual data repositories, holding e.g. scientific records or simulation results, often hold very valuable data in complex structures. The user interaction elements which are possibly contained in objects pose a whole new level of difficulties. All these specific characteristics of digital objects have to be considered and make the preservation process a difficult one.

Amongst the many strategies developed to preserve digital

JCDL'07, June 18–23, 2007, Vancouver, British Columbia, Canada.

objects and keep them accessible in the long run, migration and emulation are the most prominent ones.

Migration is the method of repeated conversion of files or objects. A file is converted to either a more current version of its own file format, or to another, which is easier to handle and access. A good example of migration to an easier preservable format is the recently adopted PDF/A standard [10]. It implements a subset of the PDF standard and is especially well-suited for long-time preservation due to its omitting of, for instance, embedded scripts. Other examples would be the conversion from Microsoft Word to RTF et vice versa.

Emulation denotes the duplication of the functionality of systems, be it software, hardware parts, or legacy computer systems as a whole, needed to display, access, or edit a certain document. In the preservation context, this most often means emulating a certain (version of) a software system needed to access a file in an outdated version or format.

The applicability of both strategies is highly problematic and context dependent and their success very specific to different settings [22]. Every scenario has its own requirements and problems, calling for different solutions to the problem at hand. Preservation strategies and specific software tools for emulation or migration must always be chosen according to requirements of individual institutions. In the case of digital libraries, for instance, migration or emulation often has to be performed for thousands of files. It is therefore a highly complex task of reaching a decision on which preservation strategy to follow. Many aspects need to be taken into account, some of which have the potential to significantly influence the financial expenditures like personnel or hardware costs. This is exactly where preservation planning fits in. Whenever preservation decisions, which are usually made by highly skilled and trained individuals, have to be made, it is utterly important to provide assistance in the process. This aspect becomes even more important when less skilled or trained staff is concerned with preservation planning. One possibility to support such decision processes are software tools which guarantee full traceability and documentation of all elements influencing the final decision. Preservation planning also means to take into account unavoidable losses that will, up to a certain extent, always be part of preservation processes, be it loss of document characteristics during migration or loss of certain ways of user interaction in emulation scenarios.

Preservation planning has been identified to be a vital aspect of the digital preservation process as a whole in the context of archival standards as well as the 'Preservation and Long-Term Access via Networked Services' EU project (PLANETS)[1]. It also forms a core functional entity in the OAIS model, the ISO-adopted Reference Model for an Open Archival Information System, which is a common reference model for archives [9].

In this paper, we will describe in detail the workflow for evaluating and selecting digital preservation solutions following the principles of the *PLANETS Preservation Planning approach*. We will describe a framework to support the acquisition and documentation of various requirements arising in the context of preservation planning. Besides, guidance is provided for institutions having less expertise in the area of digital preservation and its challenges to identify core requirements that any solution should fulfil in a given setting. The stakeholders need to precisely specify and document the goals and requirements for the envisaged digital preservation solution. Furthermore, a structured model for repeatable experiments is also needed as basis for informed decisions and hence another vital aspect of the *PLANETS Preservation Planning approach*. Once alternative preservation paths are specified and experiments are performed, Utility Analysis and its ability to integrate inhomogeneous criteria sets is used to evaluate different strategies. The strengths of Utility Analysis lie in the definition of objectives and the clear evaluation and comparison metric, which make the ranking of alternatives possible and intuitive. A first model for planning preservation solutions, on which this work is based on, is introduced in [15].

We describe a set of case studies demonstrating the feasibility of the proposed approach. Specifically, we report on three case studies from different domains covering a wide range of requirements: web archive collections, collections of scientific publications, and electronic multimedia art.

The remainder of this paper is organised as follows: Section 2 provides pointers to related initiatives and gives an overview of work previously done in this area. After that, Section 3 shortly describes the Open Archival Information System (OAIS) focusing on the role of preservation planning. Further, we give an overview of the principles of Preservation Planning and the approach pursued in PLANETS in Section 4, which also provides a detailed description of the workflow. We then report on a set of case studies in Section 5. Finally we draw conclusions, summarise lessons learned as well as give an outlook on future work in Section 6.

2. RELATED WORK

An increasing amount of cultural heritage material, legal, and scientific information is born-digital or only available in digital form. The ability of accessing and using the digital information will usually depend on a particular version of a programm on a specific computer platform. The heterogeneity and the complexity of digital formats make the preservation of the information a difficult task. At the moment libraries, archive and scientific institutions are primarily dealing with the challenge of long term preservation. Other institutions such as government agencies, large industries, SMEs and also private users, who have steadily growing amounts of legally or personally important data, are increasingly facing this problem. This results in the creation of a number of large scale initiatives integrating digital preservation capabilities in digital repository systems [18, 21]

In December 2000, the U.S. Congress appropriated US $ 99.8 million to establish the National Digital Information Infrastructure and Preservation Programm (NDIIPP)[2]. Led by the Library of Congress, the collaborative research programm is funding research in different aspects of digital preservation, including collection practices, risk analyses, legal and policy issues, and technology.

In Europe two new research projects for digital preservation started in 2006. CASPAR[3] focuses on cultural and scientific resources. The project engages the implementation

[1] http://www.planets-project.eu

[2] http://www.digitalpreservation.gov

[3] http://www.casparpreserves.eu

and extension of the OAIS reference model [9]. PLANETS[4] includes partners from library, archival and research backgrounds with a research focus on the integration of methodologies, tools and services for preservation characterisation, action, and planning. These services will be integrated to form an interoperable framework. Based on this framework the PLANETS Testbed will allow an evaluation of preservation strategies, tools, and services.

The PANIC project [7, 8] addresses the challenges of integrating and leveraging existing tools and services, and thus assists organisations in dynamically discovering possible preservation strategies. It relies on Web Services to offer preservation software, using OWL to provide semantic descriptions.

Other important projects in the field of digital preservation are the PADI project [13] from the National Library of Australia and DCC[5]. PADI identifies and promotes information of relevant activities in the field of digital preservation. Both PADI and DCC provide guidelines for preservation endeavours.

Digital Preservation Europe (DPE)[6] is a Coordinated Action of the EU aiming at better collaboration and synergies between existing preservation initiatives across Europe.

In order to support the evaluation of experiments, a number of tools and services are developed. For example, the National Library of New Zealand Metadata Extraction Tool[7] extracts preservation metadata from a range of file formats. JHove [3], developed by JSTOR and the Harvard University Library enables the identification and characterisation of digital objects. These tools can be used to analyse files of migration experiments. File format repositories, such as PRONOM [14] may be used to identify specific characteristics of digital objects at hand, helping in the definition of the preservation requirements.

The approach presented in this paper has its focus on the evaluation of preservation strategies. Therefore, the elicitation and documentation of the preservation requirements (objectives), as well as running and evaluating experiments in a structured way are required.

Over the last years a lot of effort was spent on defining, improving, and evaluating preservation strategies. A good overview of the preservation of digital heritage and preservation strategies is provided by the companion document to the UNESCO charter for the preservation of the digital heritage [22].

Research on technical preservation issues is focused on two dominant strategies, namely migration and emulation. The Council of Library and Information Resources (CLIR) presented different kinds of risks for a migration project [11]. Migration requires the repeated conversion of a digital object into more stable or current file formats, such as e.g. converting a Microsoft WORD97 document into the current Office 2007 format (within format-family migration) or converting it, e.g. to Adobe PDF/A, a simple ASCII/UNICODE text file, a screenshot image, or others. Each of this incurs certain risks and preserves only a certain fraction of the characteristics of any digital document. Conversion to PDF, for example, changes the look-and-feel as well as the behaviour of the document, some fonts may not be available on a given future computer system and are not always correctly embedded, and edit history and other metadata are likely to be lost. Screenshots may preserve the look of a document, losing the machine-readable content, i.e. the text, whereas a conversion to a text file preserves the content but loses image information, macro interactivity and others. Even migrations within the same format family may incur unwanted and unspecified changes. Still, the number of tools as well as the ease of applying this strategy make it a very promising candidate - albeit increasing the difficulty in finding an optimal solution, minimising the impact of various migration steps.

Emulation, the second important preservation strategy aims at providing programs that mimic a certain environment, e.g. the emulation of a certain processor type or emulating the features of a certain operating system. For example, to run Microsoft WORD on a Linux operating system using the WINE windows 'emulator', WINE implements a compatibility layer for the operation system. Jeff Rothenberg together with CLIR [17] envisions a framework of an ideal preservation surrounding for emulation. In order to make emulation usable in practice, several projects developed it further. One of them is the CAMILEON project with the BCC Domesday project [12]. Large scale emulation of archived office documents is shown in [16]. The Digital Asset Preservation Tool, an implementation of the Universal Virtual Computer presents an new strategy for digital preservation [4], currently the file formats JPEG and GIF87a are supported.

The challenge of preserving born-digital multimedia art, which is inherently interactive, virtual, and temporary, has been an actively discussed topic over the last years. In 2004, the ERPANET projcet organised a workshop [2] on archiving and preservation of born-digital art. Preserving the inherent complexities of interactive multimedia is a very difficult task, particularly because formats used in multimedia art are ephemeral and unstable. It also poses a paradoxon between the transformation necessary to keep the work accessible, and desired authenticity of each piece of art. Depocas [1] argues that efforts to preserve born-digital media art have to be based on structured documentation. Hunter [6] discusses a case study searching for optimum preservation strategies for selected complex mixed-media objects.

3. THE OAIS REFERENCE MODEL

The Reference Model for an Open Archival Information System (OAIS) was published 2002 by the Consultative Committee for Space Data Systems (CCSDS). ISO 14721:2003 [9] defines an OAIS as

> ... an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community.

The OAIS model further

> ... provides a framework for describing and comparing different long term preservation strategies and techniques.

The left hand side of Figure 1 shows the main functional components of the model. When a producer, i.e. a provider

[4] http://www.planets-project.eu
[5] http://www.dcc.ac.uk
[6] http://www.digitalpreservationeurope.eu
[7] http://www.natlib.govt.nz/en/whatsnew/4initiatives.html\#extraction

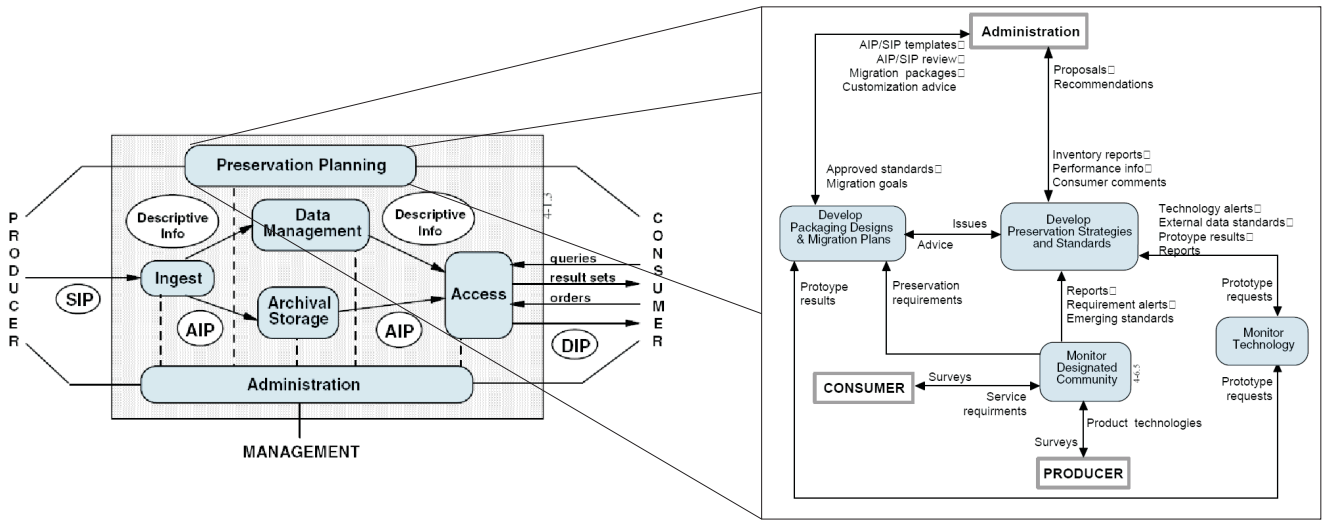

Figure 1: Functional entities of the OAIS reference model

of content, submits a digital object to the system, it has to be packaged together with required metadata as a Submission Information Package (SIP). The Ingest module provides the services and functions to accept SIPs from Producers. It further performs quality assurance and generates an Archival Information Package (AIP) complying with the archive's standards. Ingest also extracts descriptive information from the AIPs and coordinates updates to Archival Storage and Data Management.

Archival Storage stores, maintains and retrieves AIPs, while Data Management populates, maintains and accesses descriptive information about archived objects as well as administrative data. Every action inside the archive that affects the object is added to the metadata of the AIP.

The Access component is responsible for supporting consumers, i.e. users looking for content, in finding, requesting and receiving information stored in the system. Access functions include access control, request coordination, response generation in the form of Dissemination Information Packages (DIPs) and delivery of the responses to consumers.

The Preservation Planning entity monitors the environment and provides recommendations to ensure the long-term accessibility of the stored information. This includes monitoring of the technology and designated community and evaluation of the archive and periodical recommendations on archival updates for migration. A central component is the development of preservation strategies and standards as well as packaging designs and plans.

The right hand side of Figure 1 shows in detail the functional entries of preservation planning. The 'Monitor Designated Community' function and 'Monitor Technology' functions report changes of service requirements, technologies and standards technologies and designated communities to the 'Develop Preservation Strategies and Standards' function. In response to these reports the 'Develop Preservation Strategies and Standards' function starts to evaluate and develop preservation strategies and standards to ensure accessibility and usability of the current archive holdings and for new submissions. These recommendations are sent to the Administration. The *PLANETS Preservation Planning approach* presented in this paper supports the evaluation of preservation strategies and production of well-documented, accountable recommendations on which strategy to follow. The PLANETS preservation workflow covers the OAIS 'Develop Preservation Strategies and Standards' function and can be easily integrated into existing archival environments.

4. PRESERVATION PLANNING

4.1 Overview

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.

Preservation Planning, i.e. evaluating preservation strategies and choosing the most appropriate strategy, has turned into a crucial decision process, depending on both object characteristics as well as institutional requirements. The selection of the preservation strategy and tools is often the most difficult part in digital preservation endeavours; technical as well as process and financial aspects of a preservation strategy form the basis for the decision on which preservation strategy to adopt. The area of Preservation Planning has therefore attracted much interest in recent years.

In the last years two frameworks were created in parallel for supporting the establishment of a digital preservation solutions. The Utility Analysis approach developed at the Vienna University of Technology [15] and the Dutch testbed designed by the National Archive of the Netherlands [5]. The advantages of these two were integrated and form the basis for the *DELOS Preservation Testbed* [20]. The strengths of the Utility Analysis is the clear hierarchical structuring of the preservation objectives, which documents the requirements and the goals for an optimal preservation solution. Its second advantage is the numerical evaluation of the objectives, allowing a direct mathematical comparison and ranking of the alternative solutions. The strengths of the Dutch testbed are the detailed definition of the environment and the standardised experiment procedure.

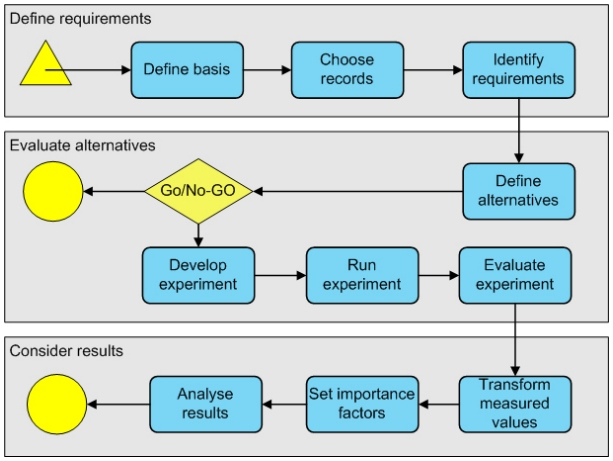

Figure 2: Overview of PLANETS Preservation Planning workflow

In the PLANETS project, the *DELOS Preservation Testbed* forms the basis for the Preservation Planning approach. The workflow was refined based on practical experience and feedback from the user community. The result of this process is the here presented PLANETS Preservation Planning approach.

4.2 Workflow

Figure 2 provides an overview of the preservation planning workflow.

The 3-phase process, consisting of a total of 11 steps, starts with defining the preservation scenario, choosing sample records for experiments, and identifying the requirements and goals for the preservation scenario.

The second part of the process consists of the definition and evaluation of potential preservation alternatives. Therefore, alternatives are identified, including technical settings and required resources for running the experiments. The Go/No-Go-Decision enforces a review of the work in the previous steps. The experiments are set up and run. The last step of the second phase is the evaluation of the experimental outcomes against the requirements and goals defined in the first phase.

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. After this consideration a clear and well argumented accountable recommendation for one of the alternatives can be made.

The detailed workflow as shown in Figure 2 is described below.

1. **Define Basis**
   In the first step the preservation scenario is described in a semi-structured way including the collection to be considered. The information about the collection consists of types of records or files, their numbers and legal issues. Moreover, the environment is described in which the preservation process takes place including institutional policies for preservation.

2. **Choose Records**
   This step selects sample records representing the variety of object characteristics of the considered collection. These samples are later used for evaluating the preservation alternatives.

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 into a tree structure.

   While the resulting trees usually differ according to specific 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 when 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* denote the technical foundations of a digital record, the context, interrelationships and metadata.
   - *Process characteristics* refer to 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 into technical and personnel costs, as well as start-up and operational expenditures.

   The objective tree is usually created in a workshop setting with experts from different domains contributing to the requirements gathering process. The tree is independent from the preservation approach, it documents the individual preservation requirements of an institution for a given collection of objects. Typical trees may contain from 50 up to several hundred objectives, usually organised in four to six hierarchy levels.

   Measurement units are assigned to each leaf in the objective tree. 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 acceptance within different communities, etc.).

4. **Define Alternatives**
   Different preservation solutions, such as different migration tools or emulators, are selected. An extensive description of each alternative including its software environment and parameters ensures a clear understanding. For each alternative defined, the amount of work, time and money required for running experiments are estimated.

5. **Go/No-Go**
   This step considers the defined requirements, the alternatives and estimated resources to determine if the

proposed alternatives are feasible. The result is a decision for continuing the evaluation process or a justification of the abandonment or postponement of certain alternatives.

6. **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.

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

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

9. **Transform Measured Values**
   The measurements taken in the experiments might all be scaled differently. 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.

10. **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 scenario.

11. **Analyse Results**

    In this step, the performance measures for the individual objectives are aggregated to one single comparable value for each alternative. Usually, the measured performance values as transformed by the transformation tables are multiplied with the weighting factor. These values are summed up. A range of other aggregation techniques have been implemented in the system, resulting in slightly more pronounced final ranking values for the alternatives. Currently the following methods are available:

    - *Weighted Sum* – The comparable values are multiplied with their weights. These values are summed up to a single comparable value per alternative. The sum method offers a final ranking on an rational scale. Leaf values that score zero (drop-out value) have no decisive effect on the final root value.
    - *Multiplication* – Here, the first step is to multiply the comparable value per leaf with the weight of that leaf. The results are then multiplied throughout the tree for the whole alternative. The final ranking is based on a rational scale. The multiplication method highlights alternatives with drop out values, as these alternatives with leaf values zero have a final root value of zero.
    - *Sum of Priority* – This method belongs to the family of frequency-advantage-rules. Two or more alternatives are compared based on an ordinal scale. Each alternative's leaf value is compared to the leaf values of the other alternatives.

      Each alternative's leaf value is computed as the number of alternatives with a lower value, e.g. if three leaves are compared, the one having the highest value is set to two, the one having the second highest value is set to one. In the case of equally scored values, the leaf value is set to zero. It is therefore not possible to make statements about the distance between the leaf values of the alternatives.
    - *Austin Slight* – This method is very similar to the Sum of Priority, but for equal scores the leaf value is increased by 0.5.

    We thus obtain aggregated performance values for every part of the objective tree for each alternative, including an overall performance value at the root level.

    A first ranking of the alternatives can be done based on the final root values per alternative. This ranking forms the basis for a documented and accountable decision for the selection of a specific solution to the given preservation challenge based on the requirements specified.

    In addition to the ranking, Sensitivity Analysis may be performed by analysing, for example, the stability of the ranking with respect to minor changes in the weighting of the individual objectives, or to minor changes in performance. This Sensitivity Analysis results in a stability value for each alternative and objective, which may further influence the final decision.

The result of this preservation planning process is a concise, objective, and well-documented ranked list of the various alternative solutions for a given preservation task considering institution-specific requirements. By providing both overall as well as detailed performance measures, stemming from a standardised and repeatable experimental setting, it forms the basis for sound and accountable decisions on which solution to implement.

5. CASE STUDIES

To evaluate the viability and the benefits of the presented approach, a series of case studies were performed with different partner institutions, ranging from national libraries and archives to multimedia museums and research institutions. Specifically, we report on five detailed case studies building upon the initial four case studies performed as part of the DELOS project [20]. These are:

- two web archives, one coming from a library context, the other one from an archiving institution,
- two collections of electronic publications with scientific provenance, coming from three European national libraries, and
- a large collection of born-digital multimedia art.

5.1 Web Archive Collections

In a joint requirements workshop, The British Library and The National Archive of the UK (TNA) each defined their requirements for the collections resulting from their web archiving activities.

The resulting objective trees show the different focus and background the two institutions have. While in a library context strong emphasis is placed on the user experience with the website, the archive needs to concentrate on risk assessment and technical characteristics.

Figure 3 shows a sub-branch of the tree constructed by TNA describing requirements on the technical characteristics of a preservation strategy. The outermost leaves of the nodes describe the measurement units assigned to the leaf objectives. In this case, the measurement units are described on ordinal scales, the values will have to be assigned manually by an expert.

In step 9 of the workflow, these units are transformed to a uniform scale. For example, *openness of documentation* could be mapped to 5, 3, and 1 (for *Standard*, *Open*, and *Proprietary*, respectively). Alternatively, if an open documentation is an essential criterion, the institution can manifest this by assigning a 'not-acceptable' value of zero to the value *proprietary*.

The objectives in the depicted sub-tree primarily deal with the risks that the collection is facing. For example, tool support for a file format is quantified by the number of tools that are currently supporting it. If this number is low, the risk of the file format becoming obsolete will probably be high. Related to this, the backwards compatibility of file formats can be seen as an indicator of stability.

These criteria and scales support the risk assessment for each potential preservation strategy by modelling the risks in a quantifiable way. Despite the manual assignment of measurement values, participants agreed that the rating these scales provide is far more useful and objective than an undocumented, intuitive decision.

5.2 Collections of scientific publications

This series of two case studies was conducted with the Austrian National Library (ONB) and the Royal Library of the Netherlands (KB). Both have to preserve scientific publications provided in formats ranging from MS Word and older word processing formats to current PDF files. The ONB will have the future obligation to collect and preserve electronic theses and dissertations from Austrian Universities. To fulfil its obligation, the ONB needed a first evaluation of possible preservation strategies for these documents according to their specific requirements. The KB in turn is responsible for preserving a collection of scientific documents from 18 scientific institutions in the Netherlands.

The resulting objective trees showed many similarities, but differed in some aspects very specific to each institution, such as requirements coming from the technical environment of the KB. An automated migration process in the KB has to run in parallel with other processes on a central server and thus needs to be configurable for load balancing to be able to limit the workload the process consumes to a certain threshold.

Another example are metadata of the objects. In the ONB the metadata of documents are held by the document management system. So it is not important to preserve the embedded metadata of an object, but to provide a documentation about changes of an object for the document management system. This is different with KB where the metadata are partly contained in the object. In this case the embedded metadata have to be preserved and enriched.

Table 1: Overall scores of the alternative strategies considered in the ONB case study

Nr.	Alternative	Total Score	
		Sum	Multiplication
1	PDF/A	4.52	4.31
2	TIFF	4.26	3.96
3	EPS	4.22	3.91
4	JPEG	4.17	3.77
5	RTF (Adobe)	3.43	0.0
6	RTF (ConvertDoc)	3.38	0.0
7	TXT	3.28	0.0

Table 1 provides the resulting ranking of some alternative migration strategies considered by the ONB. Only the root values of the Sum and Multiplication aggregations are shown.

All experiments were executed on Windows XP professional on a sample set of five master theses. The results 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 migration to 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 migration to TIFF, EPS and JPEG show very good appearance, but have weaknesses regarding criteria such as 'content machine readable'.

The aggregation method 'Multiplication' shows that the alternatives RTF (Adobe), RTF (ConvertDoc) and TXT failed to preserve essential characteristics and to fulfil the minimum requirements in at least one objective.

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 almost equal score. The migration to plain text format fails to preserve important artefacts like tables and figures as well as appearance characteristics like font types and sizes.

5.3 Electronic Multimedia Art

The Ars Electronica Center (AEC) in Linz, Austria[8] has been collecting electronic art in digital form since the early nineties. The AEC holds more than 25.000 CDs containing multimedia and interactive art in different formats like long-obsolete presentation file formats with interactive visuals, audio and video content. The Ludwig Boltzmann Institute[9] is currently evaluating alternative strategies to not only preserve these pieces of art over the long term, but also make them accessible in a satisfying form on the web.

[8] `http://www.aec.at/en/center`

[9] `http://media.lbg.ac.at/en/institution.php?iMenuID=1`

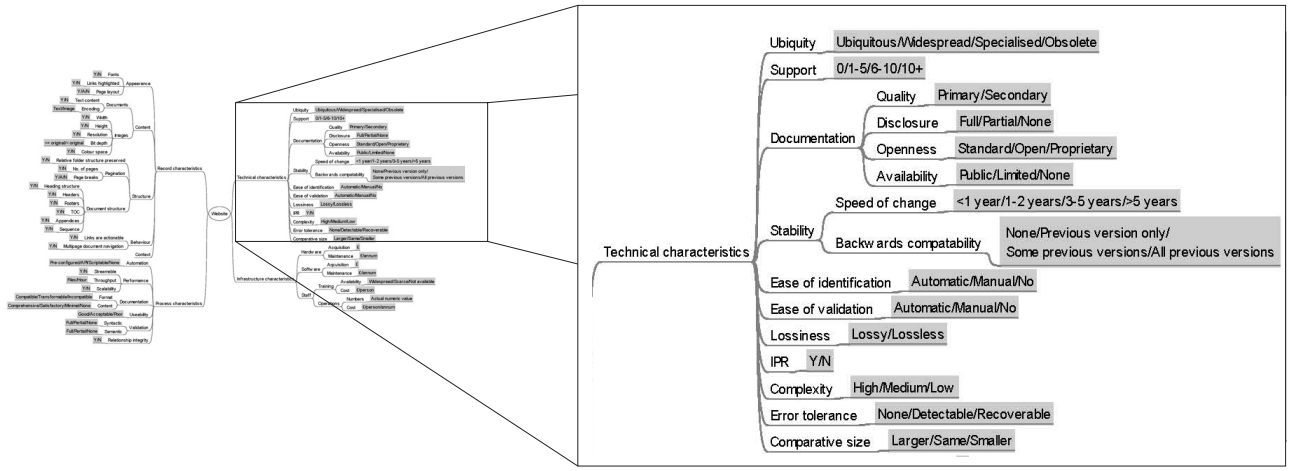

Figure 3: Technical characteristics for preserving a web archive as identified by TNA

The case study focused on requirements for a sub-collection containing interactive presentations in file formats such as Asymetrix Compel. More specifically, we concentrated on the requirements for the 'documentation' of the art objects that has to be created in order to enable long-term preservation and interactive access over the web.

The resulting objective tree is strikingly different from those that arose in other settings such as library and archival institutions as it shows a strong focus on the navigation behaviour of the artwork and the appearance of animations.

An example for a specific characteristic of multimedia art is the speed of execution. Migrating a piece of art or emulating the original platform for which it was created might alter this speed. The deviation can be measured in percent.

Further case studies that were conducted during the last years as part of DELOS [19] 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 produced 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 metadata 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 to 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 the annual electronic journal of differential equations

  Within the project of supra-regional literature supply in Germany the State and University Library Göttingen holds the collection of 'Electronic Journal of Differential Equations'. The SUB is committed to preserve 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.

The workshops held to define the objectives of the preservation endeavour usually followed the same pattern. Participants with different backgrounds (usually technical and managerial/curators) were present, resulting in groups of about three to seven people. After defining the basic settings, a brainstorming session helped to elicit as many different objectives as possible. These were then reduced and structured, forming a basic objective tree. During this process, usually numerous further objectives were identified. Specifically with regard to technical characteristics, in some cases the discussion threatened to lead to a full-fledged listing of all metadata embedded in a digital object, or a list

of highly specialised characteristics inherent in a specific file format. It was of vital importance to maintain the balance between the necessary level of detail and the actual requirements, focusing on the actual needs of the preservation process and the intended utilisation of the objects.

The assignment of measurable units to each of the leaf objectives is a very important step. In principle, the proposed method allows subjective evaluations. However, apart from the fact that these require manual evaluation during the experiment phase, they also do not offer themselves for an objective evaluation, forming the basis for accountable decisions. Thus, wherever possible, specific objective measures need to be specified. This sometimes requires a revision of the objectives identified, either re-formulation or further refinement. Participants sometimes have difficulty in quantifying characteristics that are at first perceived as too elusive for objective measurements. Moderation and guidance of the group discussion as well as illustrative examples have been very helpful in reaching useful measurement scales.

Furthermore, a precise definition and labelling of the objectives is crucial to avoid ambiguities, redundancies, or misunderstandings.

Another critical issue is the assignment of the importance factors. Standard Utility Analysis would require these to be defined immediately after identifying the objectives. However, experience showed that in many cases the various tools performed virtually identical for a range of requirements. Thus, specifying the importance values after running the experiments, although not formally correct due to possibly biased decisions for or against a certain solution, turned out to be more efficient, preventing avoidable discussions. This is also due to the 'friendly setting' the case studies generally were set in. Institutions were not looking for external certification, but for the best solution for the problems they are facing. If this approach is used for external vendor certification or a bidding process, it might be advisable to assign importance factors before evaluation. The support software tools allow for a flexible redesign of the process in this case.

In some cases there were concerns relating to the influence of the weighting on the final rankings of the alternatives. This led to the development of an automated Sensitivity Analysis evaluating the impact that a variation by a certain percentage for each of the weights would have on the overall outcome. This, however, was in most cases minimal. Sometimes the order of consecutive pairs of alternatives was switched.

In general, participants were very satisfied with the process - particularly with the elicitation of the preservation requirements, which required a structured view on the problem and the needs. The experiments to actually evaluate the various tools were considered to be of minor importance.

The elicitation and definition of requirements during the brainstorming session, as well as the subsequent structuring to form an objective tree, was initially performed in a traditional manner, using staples of post-it notes on a whiteboard. 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.

The tool support was highly welcome during the various stages of running the workshops, specifically for documenting each individual step. Figure 4 shows the consecutive editing following the case studies in the planning tool.

Select	Focus	Node	Value	Value
○		▼ Website	6.4	Add Child
○	x	▼ Record characteristics	1.9	Add Child
○	x	▼ Appearance	1.2	Add Child
○		Fonts	0.4	Add Child
○		Links highlighted	0.4	Add Child
○		Page layout	0.4	Add Child
○	x	▼ Content	0.4	Add Child
○	x	▼ Documents	0.2	Add Child
○		Text content	0.1	Add Child
○		Encoding	0.1	Add Child
○	x	▼ Images	0.3	Add Child
○		Width	0.1	Add Child
○		Height	0.1	Add Child
○		Resolution	0.05	Add Child
○		Bit depth	0.05	Add Child

Figure 4: PLANETS planning tool objective tree

Most participants had a pretty good feeling on their performance and the effects they would have on the files so that they could guess the results once the objectives were clear. Generally, in most cases the experts were able to predict the outcome, but highly valued the evaluation setting, as this provided a means to document the facts, providing a basis for an accountable decision.

6. CONCLUSIONS

We presented the *PLANETS Preservation Planning approach* to support preservation planning activities. It enforces the explicit definition of requirements for preservation endeavours in terms of specific objectives. It therefore offers a standardised way of planning and evaluating preservation strategies based on a set of experiments. Evaluation is done via Utility Analysis, helping the testbed in establishing and maintaining a trusted environment for digital preservation processes. We demonstrated the applicability of the presented approach and software tool by evaluating and testing it through case studies in a variety of domains.

Of course, the ongoing development of the planning software itself will be an important aspect of future work. The overall usefulness heavily depends on the integration of a wide range of tools for file analysis and conversion. We mentioned the role of preservation planning in form of the *PLANETS Preservation Planning approach*, yet other services are being developed. Hence, the integration with services from other parts of the project like Preservation Action and Preservation Characterisation will play an important role in the further development as well, and definitely is going to be one of the strong points of the *PLANETS Preservation Planning approach*.

Moreover, continuous case studies will be performed to keep contact to users from different domains and institutions and to ensure the pursuing usability and effectiveness of the *PLANETS Preservation Planning approach*. The case studies will further be used to establish a basis of best practice models. Based on these models we want to construct a kind of recommender process. It should provide a pre-

defined objective tree depending on the type of preservation setting, or, at least, a set of building blocks therefor.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the DELOS NoE on Digital Libraries, contract 507618, and the PLANETS project, contract 033789.

7. REFERENCES

[1] Depocas, A. Digital preservation: Recording the Recoding. The Documentary Strategy. In *Ars Electronica 2001: Takeover. Who's doing the Art of Tomorrow?* (2001). `http://www.aec.at/festival2001/texte/depocas_e.html`.

[2] ERPANET. The archiving and preservation of born-digital art workshop. Briefing Paper for the ERPANET workshop on Preservation of Digital Art, 2004.

[3] Harvard University Library. Jhove - jstor/harvard object validation environment. Website, 2005. `http://hul.harvard.edu/jhove`.

[4] Hoeven, J., Van Der Diessen, R., and Van En Meer, K. Development of a universal virtual computer (UVC) for long-term preservationof digital objects. *Journal of Information Science Vol. 31 (3)* (2005), 196–208.

[5] Hofman, H., Verdegem, R., Day, M., Rauber, A., Thaller, M., and Ross, S. DELOS deliverable 6.1: Framework for testbed for digital preservationexperiments. Tech. rep., DELOS Network of Excellence, November 2004.

[6] Hunter, J., and Choudhury, S. Implementing preservation strategies for complex multimedia objects. In *The Seventh European Conference on Research and Advanced Technology for Digital Libraries (ECDL'03)* (Trondheim, August 17-22 2003), pp. 473–486.

[7] Hunter, J., and Choudhury, S. A semi-automated digital preservation system based on semantic web services. In *Proceedings of the Joint Conference on Digital Libraries(JCDL'04)* (Tucson, Arizona, USA, June 7-11 2004), ACM, pp. 269–278.

[8] Hunter, J., and Choudhury, S. Panic: an integrated approach to the preservation of composite digital objects using semantic web services. *International Journal on Digital Libraries 6*, 2 (2006), 174–183.

[9] ISO. *Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003)*, 2003.

[10] ISO. *Document management - Electronic document file format forlong-term preservation - Part 1: Use of PDF 1.4 (PDF/A) ISO/CD 19005-1*, 2004.

[11] Lawrence, G. W., Kehoe, W. R., Rieger, O. Y., H.Walters, W., and Kenney, A. R. *Risk Management of Digital Information: A File Format Investigation*. Council on Library and Information Resources, Washington D.C., USA, June 2000.

[12] Mellor, P. Camileon: Emulation and BBC domesday. *RLG DigiNews 7*, 2 (April 2003).

[13] PADI: Preserving access to digital information, 2006. `http://www.nla.gov.au/padi`.

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

[15] Rauch, C., and Rauber, A. Preserving digital media: Towards a preservation solution evaluation metric. In *Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL'04)* (Shanghai, P.R. China, December 13-17 2004), Springer, pp. 203–212.

[16] Reichherzer, T., and Brown, G. Quantifying software requirements for supporting archived office documents using emulation. In *Proceedings of the 6th Joint Conference on Digital libraries (JCDL'06)* (New York, NY, USA, June 11-15 2006), ACM Press, pp. 86–94.

[17] Rothenberg, J. Avoiding technological quicksand: Finding a viable technical foundationfor digital preservation, January 1999. `http://www.clir.org/pubs/reports/rothenberg/contents.html`.

[18] Smith, M. Eternal bits: How can we preserve digital files and save our collective memory? *IEEE Spectrum 42*, 7 (July 2005).

[19] Strodl, S., Rauber, A., Hofman, H., Bogaarts, J., Verdegem, R., Kaiser, M., Bettelli, E. N., Neuroth, H., Strathmann, S., Debole, F., and Amato, G. DELOS deliverable 6.6: Delos digital preservation testbed for testing and evaluating digital preservation solutions. Tech. rep., DELOS Network of Excellence, July 2006.

[20] Strodl, S., Rauber, A., Rauch, C., Hofman, H., FrancaDebole, and Amato, G. The DELOS testbed for choosing a digital preservation strategy. In *Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL'06)* (Kyoto, Japan, November 27-30 2006), Springer, pp. 323–332.

[21] Tansley, R., Bass, M., Stuve, D., andDaniel Chudnov, M. B., McClellan, G., and Smith, M. Managing resources and services: The dspace institutional digital repositorysystem: current functionality. In *Proceedings of the 3rd Joint Conference on Digital Libraries (JCDL'03)* (Houston, USA, May 27-31 2003), ACM, pp. 87–97.

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